EVALUATING THE POTENTIAL FOR USING SPECIES PRESENCE DATA COLLECTED BY COMMERCIAL FISHERMEN FOR SPECIES DISTRIBUTION MODELING IN THE GULF OF MAINE

by

Julia Livermore

Dr. Patrick Halpin, Advisor

Dr. Kathy Mills, Co-Advisor

May 2015

Masters project proposal submitted in partial fulfillment

of the requirements for the Master of Environmental Management

degree in the Nicholas School of the Environment

of Duke University

2015

Executive Summary

Fishermen and scientists have noted that fish species distributions are changing along the Northeast Continental Shelf. The Northeast Shelf Large Marine Ecosystem has undergone changes in temperature, stratification, and circulation patterns due to local climate variability and large scale warming trends. In response to physical changes, there have been adjustments to phytoplankton dynamics, and changes in the abundance and distribution of economically important marine fish and invertebrates, including groundfish and lobsters. Northeast Fisheries Science Center (NEFSC) semiannual trawl survey data have been modeled to demonstrate these distributional and poleward shifts in species assemblages. Many New England groundfishermen argue that management structures are slow to reflect these changes and that species distribution modeling of trawl survey data has failed to accurately characterize them, as the NEFSC trawl survey may not be frequent enough or sensitive enough to pick up on distributional changes as they occur.

Since commercial fishermen are on the water more frequently, they have requested an outlet to submit information to scientists regarding where they are encountering shifting species while fishing. Data collected by fishermen themselves may be suitable for gaining a deeper understanding of moving fish populations. In response, the Gulf of Maine Research Institute and Island Institute have proposed to create a phone application that would allow fishermen to take photographs of species they encounter, logging geographic coordinate information at the time of the photograph.

The purpose of this study is to evaluate whether data submitted though such an application would be useful to scientists in modeling distributional shifts of commercially important species. In order to determine the fitness of potential citizen science data, I created a proxy for data submitted by commercial fishermen and used it as input data in maximum entropy models to determine whether such data would be comprehensive enough or spatially distributed in a manner that would make it suitable for species distribution modeling. The following steps were carried out to evaluate proxy data quality for a single species known to be currently on the move, red hake:

- A baseline maximum entropy model was carried out using only presence data from the NEFSC trawl survey. Models were evaluated via the AUC (area under the curve) of the Receiver Operator Characteristic (ROC).
- Complete NEFSC trawl data were used to produce a general additive model (GAMs), of the probability of encountering red hake. The GAM outputs were used to determine areas of potential habitat by selecting areas with a probability of greater than the maximum value of the Youden-Index of the ROC.
- Vessel Trip Reporting (VTR) data were provided by the National Marine Fisheries Service for the haddock and pollock fisheries (both of which intersect with the selected species' current habitat). The fishing data that overlapped with the trawl survey data temporally were overlaid onto the GAM habitat models. Areas where both fishing and modeled habitat occured were then selected as regions were fishermen could potentially submit presence data.

- Random points were generated within the selected regions (using fishing density to weight the random point assignment) to create "citizen science presence points recorded by fishermen." These points then served as the input presence-only data for maximum entropy models. Pseudopresence points were generated in nine different formats: differing combinations of three sample sizes and three levels of clustering.
- Pseudopresence data were used in maximum entropy models and evaluated using the AUC test statistic.
- The baseline trawl model and proxy data models were then compared to one another to evaluate fitness.
- All data analyses were carried out twice for red hake, once using data from fall 2013 and again for spring 2014.

All "citizen science" datasets of varying sample size and spatial distribution produced models with AUC values greater than 0.75, while the baseline models using presence data from the trawl survey had lower AUC values of 0.716 (fall 2013) and 0.704 (spring 2014). These scores are acceptable for presence only models. The proxy data models highlighted areas of high probability of encountering red hake, while the trawl survey presence models indicated larger areas of moderate encounter probability.

Our results indicate that concentrated data from fishing areas can lead to underestimation of species presence probability in regions where no fishing occurred. It is crucial to note that the VTR fishing data used to generate random points from fishermen only covered the Gulf of Maine region. Thus, we would not expect the model to demonstrate red hake preference for off-shelf habitats or areas south of the fishing zones because these regions lacked input presence data.

Findings of this study suggest that presence data from fishermen *are* suitable for presence-only maximum entropy species distribution models. Citizen science data collection and further modeling should include data from fishermen in the Mid-Atlantic as well to produce models that cover the entire study region sampled by the NEFSC trawl survey.

Table of Contents

1. Introduction

The Gulf of Maine has historically been home to some of the world's most productive fisheries (Dobbs 2000), yet fishery management efforts in New England have failed to prevent overfishing and the once lucrative cod fishery has now collapsed (Greene, 2002). To complicate matters further, the Northeast Shelf marine ecosystem has undergone considerable changes on a variety of temporal scales (Friedland and Hare, 2007; Greene and Pershing, 2007; Mountain and Kane, 2011; Mills *et al,*. 2013, MERCINA, 2013). Phytoplankton dynamics have changed in response to changes in ocean temperatures, circulation patterns, and stratification caused by global climate change (Ji *et al*., 2007). Consequently, the Northeast Shelf zooplankton community composition has been altered as well (Pershing *et al*., 2005). These physical changes and adjustments to lower trophic level interactions have also led to changes in the abundance and distribution of economically important marine fish and invertebrates, including groundfish and lobsters (Refer to Figure 1; Fogarty *et al*., 2008; Nye *et al*., 2009; Mills *et al*., 2013; Baum and Worm, 2009; Hudson and Peros, 2013).

New England fisheries have transformed with respect to changes in abundance, distribution, and biological characteristics (timing of spawning and migration) of target species. Timing and location of fishing has adjusted, making many management efforts less effective (Mills *et al*., 2013). Fishermen are highly aware of the changes occurring in their fishing grounds (Singer, 2013), but they are restricted to adjust to said changes by outdated management structures (Hudson and Peros, 2013; Pinsky and Fogarty, 2012). In light of ineffective management measures and economic hardship caused by the changes occurring in the Gulf of Maine, New England fishermen have called for an opportunity to contribute data regarding species distributions in the form of

photographs and location data. They have requested that fishery managers evaluate current data collection methods and incorporate collaborative research by using fishermen's photographs of species whose assemblages are thought to be moving geographically. Groundfishermen have suggested that traditional data collection, i.e. NOAA and NEAMAP trawl surveys, may not be sensitive to such changes (Island Institute, 2013). Historical data collected by amateurs to ecology are currently being used to understand long-term changes in ecosystems, and can increase engagement of the public in ecological research and lead to new scientific insights (Miller-Rushing *et al*., 2012; Ames, 2004). In fact, citizen science data have been essential to documenting poleward range shifts for numerous taxa across the world; these data have provided some of the strongest evidence that species are responding to global climate change (Hickling *et al*., 2006; Parmesan and Yohe, 2003; Walther *et al.*, 2002).

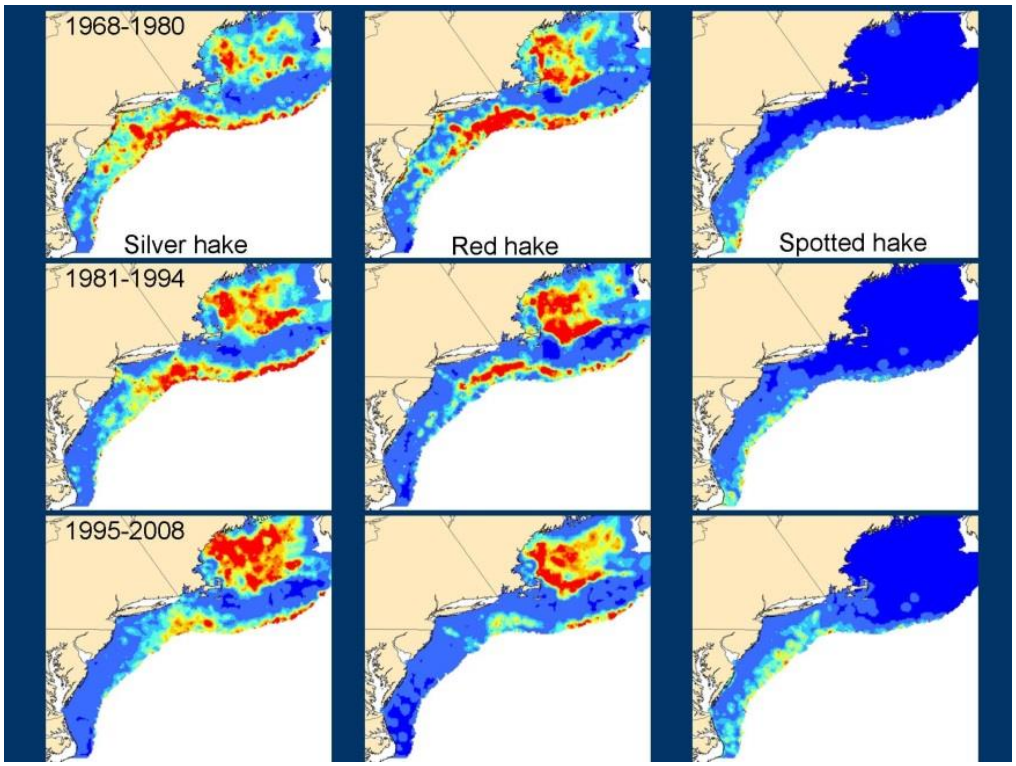


Figure 1. Distribution of silver hake, red hake and spotted hake 1968-2008. (Credit: Janet Nye, NEFSC/NOAA; http://www.nefsc.noaa.gov/press_release/2009/SciSpot/SS0916/)

In response to the fishermen's pleas for an additional data collection strategy for collaborative climate change and fisheries research, the Gulf of Maine Research Institute (GMRI) and the Island Institute have proposed to create a citizen science initiative that would allow fishermen to easily contribute their data to management efforts. The system may involve a website where data entries can be logged, but more importantly, a smartphone application that would allow fishermen to log entries while at sea. Smartphones and tablets are extremely useful tools for fishery managers because of their convenient hardware accessories including cameras and GPS. Fishery professionals are already using phone applications for public outreach efforts and as of 2013, over 56% of the US adult population has a smartphone (Gutowsky 2013). Fishermen may take photographs on their smartphone of any fish, or aquatic organism, which they feel is located in a different area than usual, and the phone's location will automatically be recorded when the photograph is captured. Thus, the dataset that results will be a presence-only dataset of species with potentially moving distributions.

While most studies involving analyses of citizen science data have to deal with data accuracy, this issue is not of concern for this particular study. Citizen scientists vary in their ability to identify species and are proven to do worse than trained biologists, even after volunteer training (Fitzpatrick *et al*., 2009). Both observer skill and inter-observer variation should be controlled for in sampling design or data analysis because they can vary widely with species (de Solla *et al*., 2005; Genet and Sargent, 2003; Lotz and Allen, 2007; Pierce and Gutzwiller, 2007; Weir *et al*., 2005). Many citizen science initiatives have dealt with this issue by introducing a data validation component where trained professionals must confirm or deny all submissions. While data validation could be implemented to verify fishermen's data, it is most likely not necessary, as

commercial fishermen are trained professionals themselves and can identify fish species very accurately. Their work is dependent on this ability.

The proposed research objective is to determine how useful the resulting dataset of presence-only data will be within the scope of fisheries management in New England. Presence-only data can be used to model species distributions by isolating the probability distribution with the maximum entropy which is closest to uniform (Valavanis *et al*., 2008). Maximum entropy modeling using presence-only data can be a useful tool in calculating the probability of encountering a species in geographic space and has been proven to effectively model species distributions when adequate data are available (Elith et *al.*, 2006). Hence, the research question is one of data quality: Will the fishermen's data be comprehensive enough or spatially distributed in a manner that will allow accurate modeling of species distributions that are on the move? If not, could the data be used to complement other data sets like the NEFSC trawl survey to produce more accurate species distribution models (SDMs)?

2. Methods

For each species, three types of models were run: (i) a general additive model to determine areas of probable species habitat, (ii) a maximum entropy model using presence data from the NOAA trawl survey, and (iii) maximum entropy models using proxy presence data. All environmental parameter data were held constant for all models during all seasons, with the exception of sea surface temperature, which was specific to the season and changed correspondingly.

2.1. Environmental data collection

Environmental/oceanographic variables were prepared on a 0.25 º latitude × 0.25 º longitude resolution grid, as this was the resolution of the sea surface temperature layers which had largest resolution of all the input variables (Figure 2). The study region covered includes the Gulf of Maine, George's Bank, Southern New England, and the mid-Atlantic Bight. Most data were publically available through the Northeast Ocean Data Portal and were compiled by members

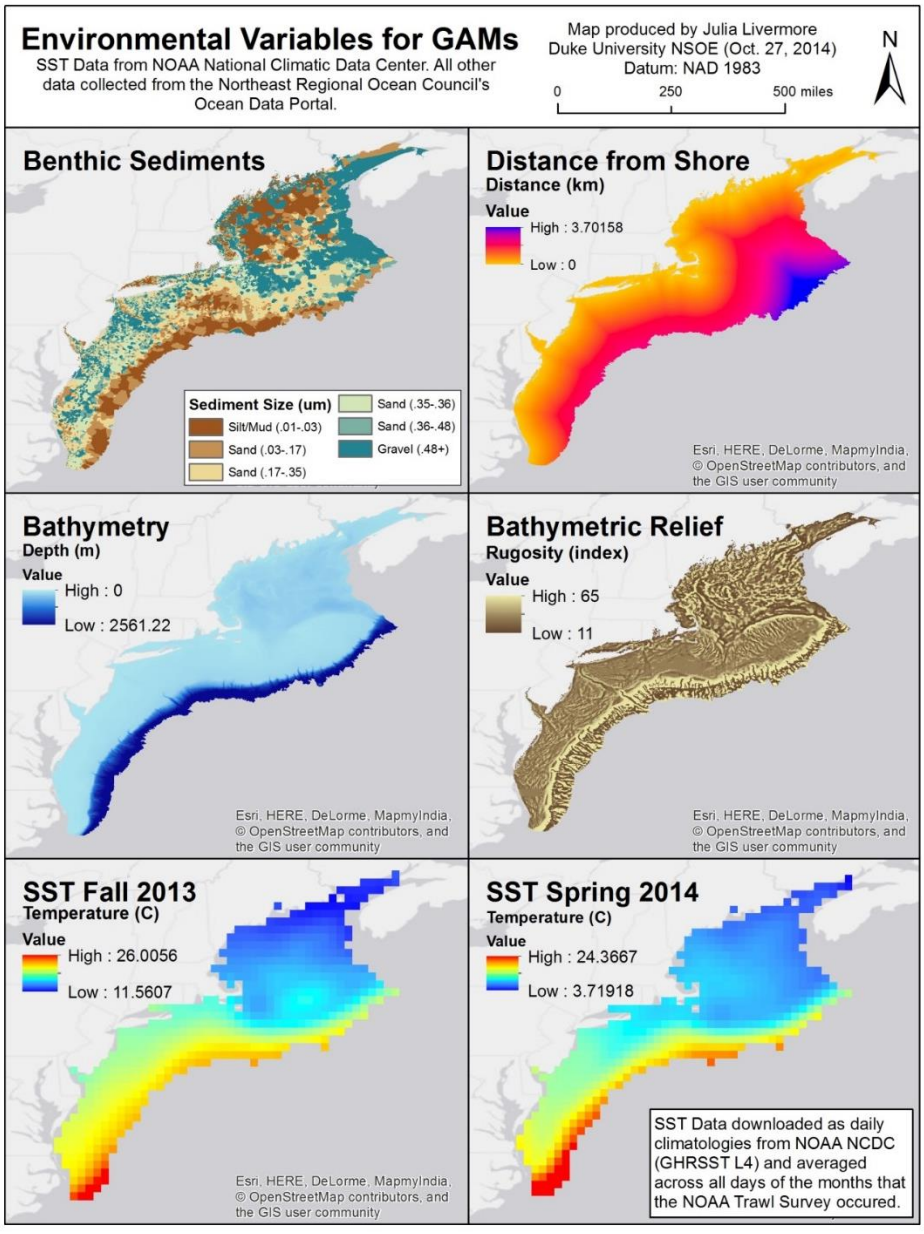


Figure 2. Environmental data used in SDMs

of the Northeast Ocean Data Group (Table 1; NROC, http://www.northeastoceandata.org/data/data-download/). Additional data included sea surface temperature data, also publically available through the National Oceanic Data Center's Group for High Resolution SST (GHRSST; NODC, http://www.nodc.noaa.gov/sog/ghrsst/accessdata.html). SST daily climatologies were downloaded through the Marine Geospatial Ecology Toolbox (MGET, www.mgel.env.dke.edu/mget) and averaged across sampling times during which the National Oceanic and Atmospheric Administration's trawl survey took place using the raster calculator function in ArcMap v10.2 (ESRI, 2013). The decision to use certain environmental variables in the SDM was dependent on data availability and biological importance. As previously stated, all SDMs used the same environmental data, though the inclusion of each environmental variable was dependent on the individual model.

Table 1. Environmental data used in all SDMs and retrieval locations

Variable	Description	Source
Bathymetry	Minimum and maximum depth	Northeast Regional Ocean Council's Ocean Data Portal (northeastoceandata.org; http://www.northeastoceandata.org/data/data-download/)
Bathymetric Relief	Relative index of rugosity	Northeast Regional Ocean Council's Ocean Data Portal (northeastoceandata.org; http://www.northeastoceandata.org/data/data-download/)
Benthic sediments	Sediment grain size and type	Northeast Regional Ocean Council's Ocean Data Portal (northeastoceandata.org; http://www.northeastoceandata.org/data/data-download/)
Distance from coast (km)	Nearest distance of each cell to the coast	Raster created by calculating Euclidian distance from shoreline (obtained through the Northeast Regional Ocean Council's Ocean Data Portal; http://www.northeastoceandata.org/data/data-download/)
Sea surface temperature (SST)	Mean sea surface temperature (degrees C) during seasonal NOAA trawl surveys)	NOAA National Climatic Data Center (downloaded as daily climatologies from GHRSST L4 for all days that the NOAA trawl survey occurred and then averaged together to form seasonal rasters used in models; this was done for both the fall 2013 and spring 2014 trawl survey time periods; http://www.nodc.noaa.gov/sog/ghrsst/accessdata.html)

2.2. Study taxon and input data

Red hake (*Urophycis chuss*), scup (*Stenotomus chrysops*), and black sea bass (*Centropristis striata*) distributions are all thought to be shifting poleward in response to environmental changes and fishing pressure (Nye *et al*., 2009; Bell *et al*., 2014). All three species are commercially and recreationally sought after. Red hake are most commonly found in the Western Gulf of Maine. They spend spring and summer months in shallow inland waters to spawn and migrate offshore in the winters. Scup are heavily fished commercially and recreationally and are known to spawn along the inner continental shelf and spend most of their adulthood along the mid and outer continental shelf. Black sea bass are a small species of grouper and can be found in inshore waters, as well as offshore in waters up to 130 m of depth. They are highly sought after by both commercial and recreational fishermen and have been overfished in the South Atlantic. All three are species known to use both inshore and offshore regions.

Fish data location data were provided by the National Marine Fisheries Service (NOAA Fisheries/NMFS) from their semiannual trawl survey carried out by the Northeast Fisheries Science Center (NEFSC; Figure 3). The trawl survey covers the Northeast United States continental shelf and occurs during the fall and spring each year; it has taken place since 1968. The survey utilizes a stratified random design and indicates the presence and absence, as well as abundance and biomass, of all species caught. Refer to Azarovitz (1981) for full data collection and sampling methods. Presence and absence values at all sampling locations (for the three target species during both the fall 2013 and spring 2014 surveys) were extracted from the database using R software.

2.3. General additive modeling and model evaluation

We used the mgcv package in R to carry out general additive models (GAMs) in order to determine each species' ecological niche or habitat preference. A logit link function was used in the model. We took an iterative approach to develop the final species-niche model for each species in order to determine which environmental variables should be included in the model to produce the best results. The overall performance of each species-niche model run was compared to other runs using the Akaike information criterion (AIC) and variable parameter significance p-values. Variables were removed and added to select the models that minimized both the AIC and p-values of included variables.

In order to measure SDM discriminative power and accuracy, we used the threshold-independent area under the curve (AUC) of the receiver operating characteristic (ROC) plot (Fielding and Bell, 1997). The ROCR package was used to carry out AUC analyses (Sing *et al.*, 2007; http://rocr.bioinf.mpi-sb.mpg.de). This package plots the ROC curve as the true positive rate (number of true positives/number of confirmed positive samples) versus the false positive rate (number of false positives/number of confirmed negative samples). A toolbox in ArcGIS was also developed to carry out modeling and habitat selection (refer to Appendices 1 -3). AUC is a commonly used test statistic because it permits a threshold-independent assessment of model performance. It is interpreted as the probability that an indiscriminately selected presence location is ranked above a random background location, designating the quality of location ranking with respect to suitability (Phillips *et al.*, 2006; Jones *et al.*, 2012). A random ranking has an average AUC of 0.5, while an AUC > 0.75 provides a constructive amount of discrimination between locations where a species is present and those where it is absent (Elith *et al.*, 2006). Areas of potential species habitat were identified by selecting regions with a probability of species

occurrence of or greater than the maximum value of the Youden Index (*J*, see Perkins and Schisterman, 2006).

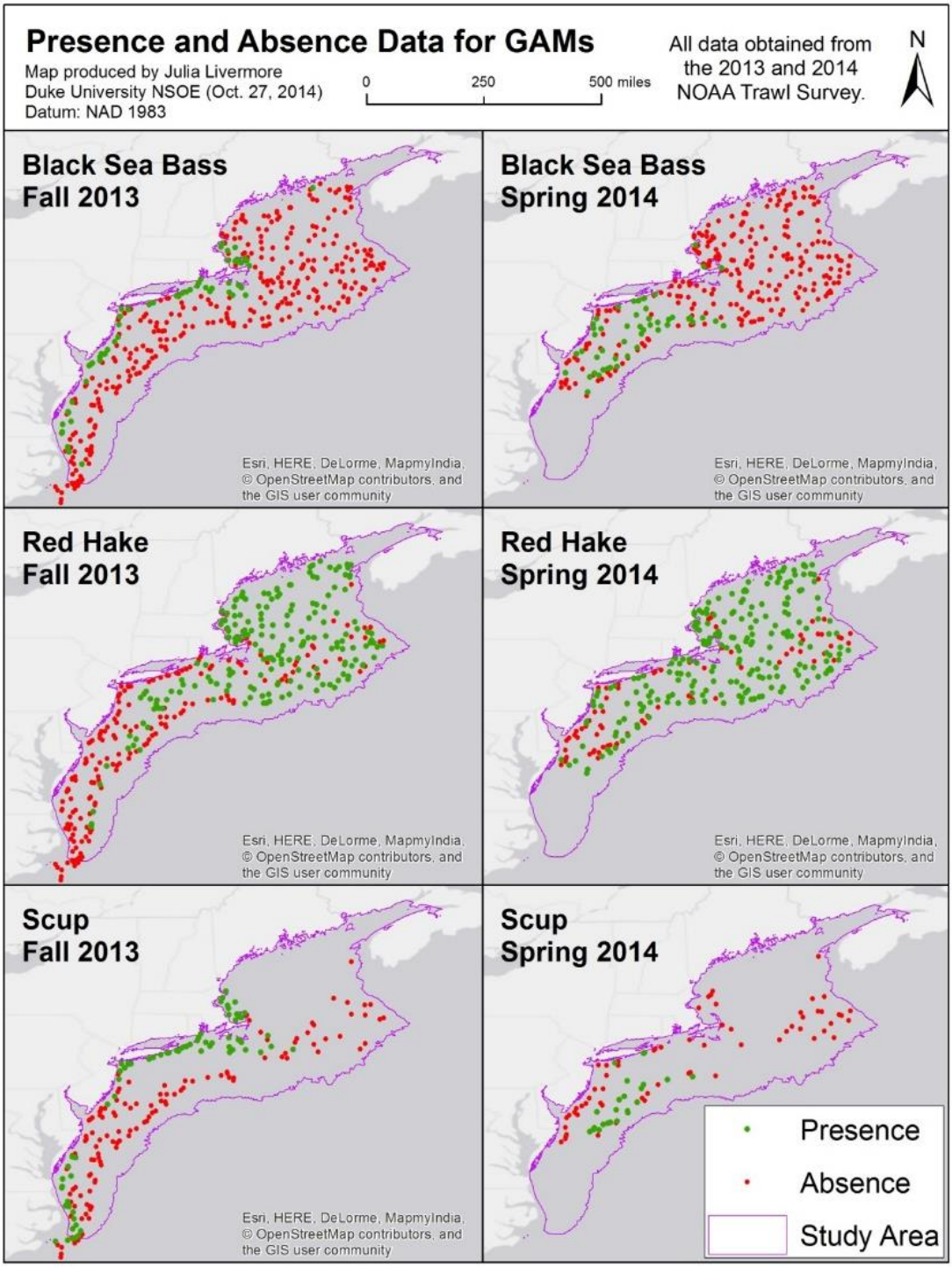


Figure 3. NEFSC trawl survey data used in general additive models and initial maximum entropy models (presence-only).

2.4. Maximum entropy models of presence data from trawl surveys

We ran Maxent version 3.3.3k (www.cs.princeton.edu/~schapire/maxent; Phillips *et al*., 2006) using the default settings: maximum number of background points = 10,000 (non-presence points); regularization multiplier = 1.000 (included to reduce over-fitting); random test percentage = 25; maximum iterations = 500; and convergence threshold = 0.00001. Maxent (Phillips *et al*., 2004) uses a multiplicative strategy to estimate the environmental co-variates that condition species presence and bases the ultimate calculation on the principle of maximum entropy. This dictates that the best approximation of an unknown distribution is the probability distribution with maximum entropy, bound by the restraints enforced by the sample of presence observations for the species (Phillips *et al*., 2006). Research has demonstrated that Maxent performs well in comparison to other conventional presence-only approaches (Elith *et al.*, 2006) and still performs well when the sample size is small (Pearson *et al.*, 2007). For this reason, Maxent is frequently used with citizen science data sets. Only presence points from the trawl survey were used as input species location data for the baseline Maxent model. Presence-only models were also evaluated using the AUC test statistic.

2.5. Proxy citizen-science data development

Haddock and pollock fishing vessel density and pounds harvested per 10 minute square of ocean were provided by the National Marine Fisheries Service from Vessel Trip Report data directly from fishermen in the Gulf of Maine (Figure 4). Dates of fishing corresponded directly to the dates that the NEFSC trawl survey occurred for the fall and spring seasons. Data were provided in a text format and converted into a raster to allow for spatial analysis (Appendix 6). The two fisheries were treated as one during further analysis; number of boats and poundage were

combined. Pounds of fish captured were divided by the number of boats in order to represent fishing effort. Using ArcGIS, the fishing effort raster was overlaid on areas of fish habitat, determined via the earlier GAMs, and only areas where both fishing and fish habitat occurred were used in further analyses. Finally, using the Geospatial Modelling Environment version 0.7.3.0 (http://www.spatialecology.com/gme/; Beyer, 2012), random points were generated using the fishing effort raster as a probability density surface. Thus, more points were generated in areas where more fishing occurs. Random points were generated with three sample sizes and three levels of clustering. Sample sizes were as follows: 10, 25, and 50 presence points. Clustering was dictated by a minimum amount of dispersion between sample points. Levels of dispersion were: 0 km (no limit on clustering), 1 km, and 3 km (clustering minimized) between points. The 3 km minimum clustering distance was selected due to the size constraint of the study area.

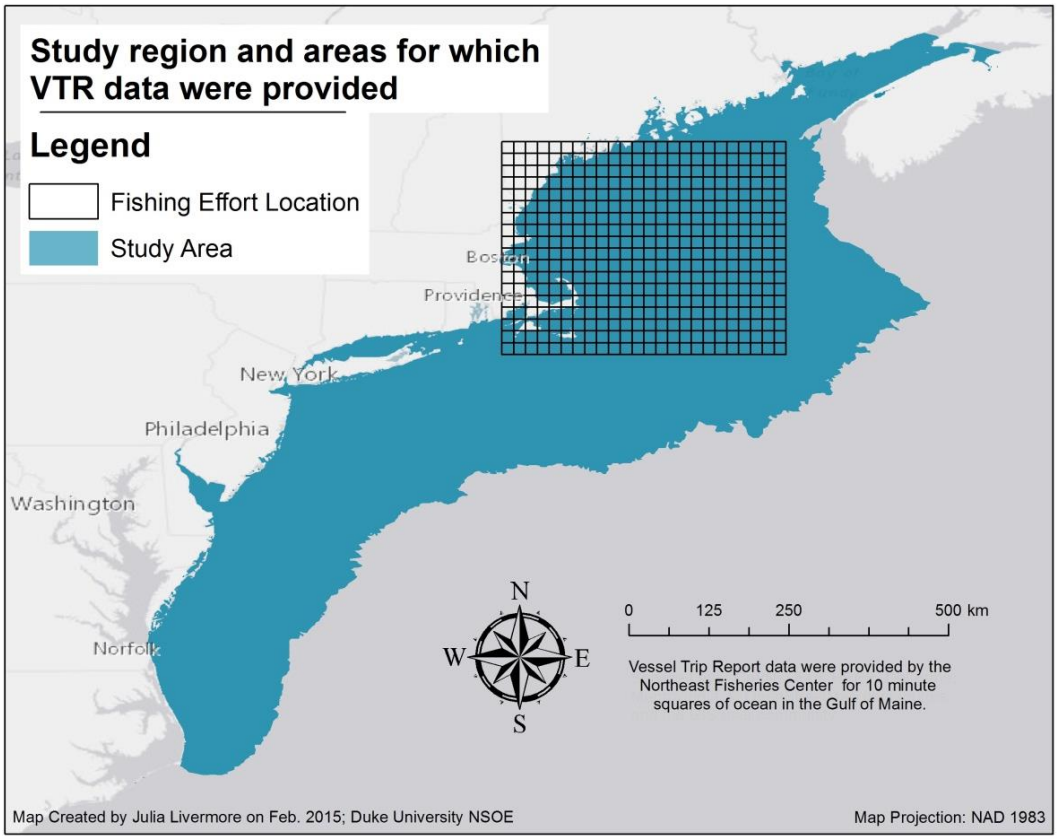


Figure 4. VTR fishing data extent provided by the NEFSC

2.6. Maximum entropy models using proxy data

In order to minimize any effects of random point generation on model outputs, we carried out 10 model runs for each Maxent model using proxy data, meaning that for each combination of sample size and level of clustering, 10 sets of unique randomly generated points were created and used as input species data in Maxent.

2.7. Independent tests of proxy model accuracy

The averages of ten replicate Maxent runs, using uniquely randomly generated input point data sets, were viewed as one model prediction and assessed further. AUC values for each individual model type (sample size and level of clustering) were averaged and analyzed in addition to the model outputs.

2.6. Model comparison

Trawl survey Maxent model outputs for each species were then compared to the averaged outputs from the proxy Maxent models. For each season and species the trawl survey model was used as the baseline and the averaged proxy models were subtracted using the raster calculator in ArcGIS. Output rasters were used to identify areas in which the SMDs were similar and areas in which they differed.

3. Results

3.1. Species selection

Complete analyses were intended to be carried out for all three target species. Due to data limitations, complete analyses were not possible for scup and black sea bass. Scup GAM performance for spring 2014 was insufficient for determining species habitat (high AIC values, low variable significance, and low deviance explained; Appendix 4), preventing any form of proxy data from being created for the species. While black sea bass produced adequate GAMs for both seasons (Appendix 5), no fishing data intersected with habitat, and therefore no pseudopresence points could be generated for the species. Thus, modeling of proxy citizen science data was only possible for red hake for fall 2013 and spring 2014. Maximum entropy models were carried out for scup and black sea bass, though they were not used for comparison (Appendices 7 and 8).

3.2. General additive models

For red hake in fall 2013, the model described 30% of the deviance (AIC = 346.4541; Figure 5). Variables included in the final spring GAM were bathymetry, bathymetric relief, and SST. The model produced and AUC of 0.854 with an accuracy of 0.778 (true positive rate/sensitivity = 0.842; false positive rate/fallout = 0.274; true negative/specificity rate = 0.726; false negative/miss rate = 0.158; refer to Table 2 for confusion matrix of values used in calculation). For red hake in spring 2014, the model only described 22.3% of the deviance (AIC = 273.8778). Variables included in the final fall GAM were bathymetry, sediments, distance to shore, and SST. The model achieved an AUC of 0.775 with an accuracy of 0.702 (true positive rate/sensitivity = 0.729; false positive rate/fallout = 0.303; true negative/specificity rate = 0.697;

false negative/miss rate = 0.271; refer to Table 3 for confusion matrix of values used in calculation). The low levels of deviance explained indicate that the models are not able to entirely define the ecological niche space. It is essential to note that false positives are not necessarily inaccuracies and can occur on an ecological basis (e.g. actual habitat may have been unoccupied during sampling; Hare *et al*., 2012). False negatives are more disconcerting (i.e. failing to identify areas of habitat that are actually used). False negatives were relatively low for both seasons of red hake modeling. Modeling indicates that red hake come inshore during warmer months and return to deep water off the continental shelf or deeper areas of the Gulf of Maine in the winter.

Table 2. Confusion matrix showing the number of test samples successfully and unsuccessfully predicted with the GAM for red hake in fall 2013

		Observed		**Total**
		Present	Absent	
Predicted	Present	154	43	197
	Absent	29	114	143
Total		183	157	340

Table 3. Confusion matrix showing the number of test samples successfully and unsuccessfully predicted with the GAM for red hake in spring 2014

		Observed		**Total**
		Present	Absent	
Predicted	Present	113	12	125
	Absent	90	64	154
Total		203	76	279

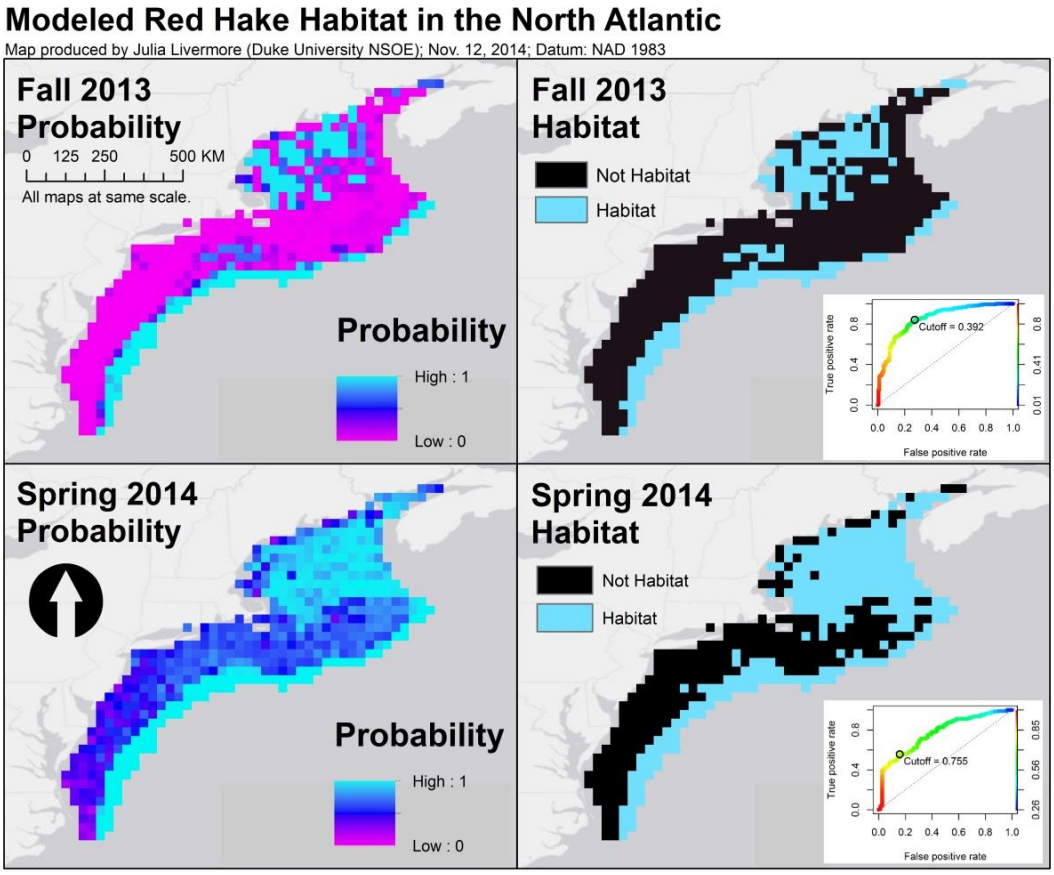


Figure 5. General additive model results from NEFSC trawl survey data. Habitat regions were differentiated from non-habitat based on the maximum Youden index value depicted in the corresponding Receiver Operator Characteristic curve plots above.

3.3. Maximum entropy models of NEFSC trawl presence data

Maxent models for red hake using only presence points from the trawl survey performed sufficiently well. While an AUC > 0.75 is preferable, a value greater than 0.7 is still moderately predictive. The fall 2013 model produced an AUC of 0.716 (Figure 6) and the spring model had an AUC of 0.704. These models were much less predictive than their GAM counterparts in terms of highlighting areas in which red hake are most likely to occur. They indicate a ubiquitous distribution across most of the study area and fail to identify the preference for offshore regions.

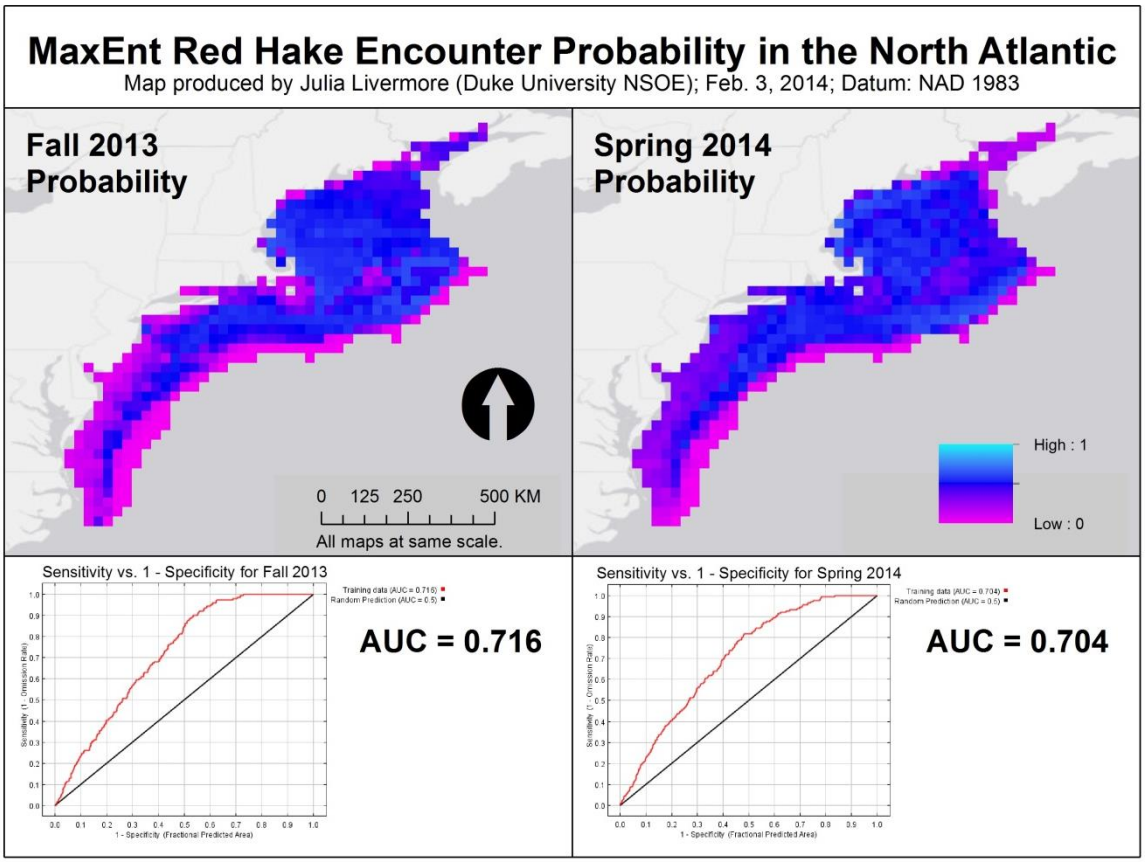


Figure 6. Maximum entropy model results from NEFSC trawl survey data (presence values only).

3.4. Maximum entropy models of proxy data

Maxent models of proxy data of all sample sizes and levels of clustering produced fairly different results from Maxent models using trawl presence data (Figures 7 and 8; refer to Appendices 9 - 26 for enlarged versions of all averaged models). All models for both seasons produced AUC test statistics greater than 0.75; their performance was significantly better than that of the trawl survey Maxent models. All fall proxy models indicated a preference for the Gulf of Maine over the Mid-Atlantic Bight, especially near George's Banks (Figure 7). The spring models highlighted similar regions but also indicated George's Bank as an area of high probability of occurrence. During both seasons, it is clear that the model increases specificity as the sample size increases, as the regions of high probability become smaller and greater in probability value. The

individual proxy models with 10 data points performed similarly but still better than the trawl presence model; the worst performing individual proxy model (a single replicate of spring 2014, N=10, 1 km dispersion between points) had an AUC test statistic of 0.687 (Tables 4 and 5). The maximum AUC test statistic was 0.956 (a single replicate of spring 2014, N=50, 1 km dispersion between points). Increasing the sample size for proxy presence datasets significantly increased the mean AUC (Figures 9 and 10) and significantly decreased the standard error of that mean (Figures 11 and 12). Mean AUC scores for sample sizes 25 and 50 were very similar for both seasons and all levels of clustering. In terms of the effects of spatial distribution of sample points on model outputs, there was no clear trend; three levels of clustering was not sufficient to demonstrate any trend.

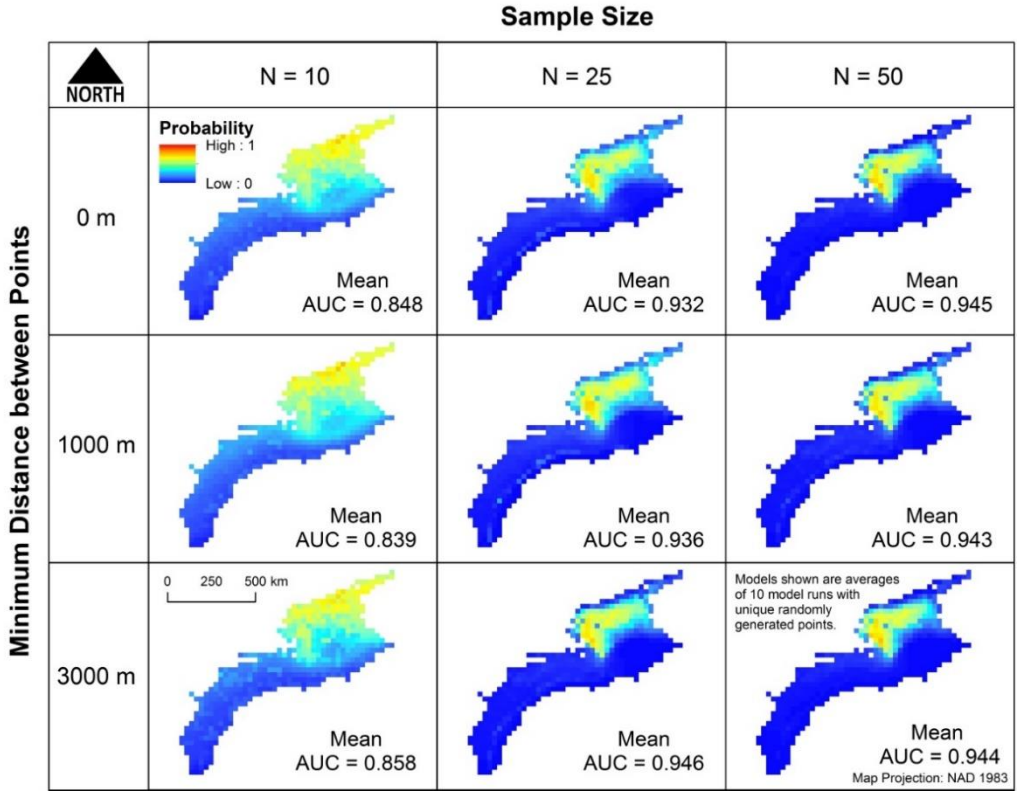


Figure 7. Final averaged maximum entropy models for red hake in fall 2013 using 10 uniquely and randomly generated input point datasets for each model type (combination of sample size and dispersion between points).

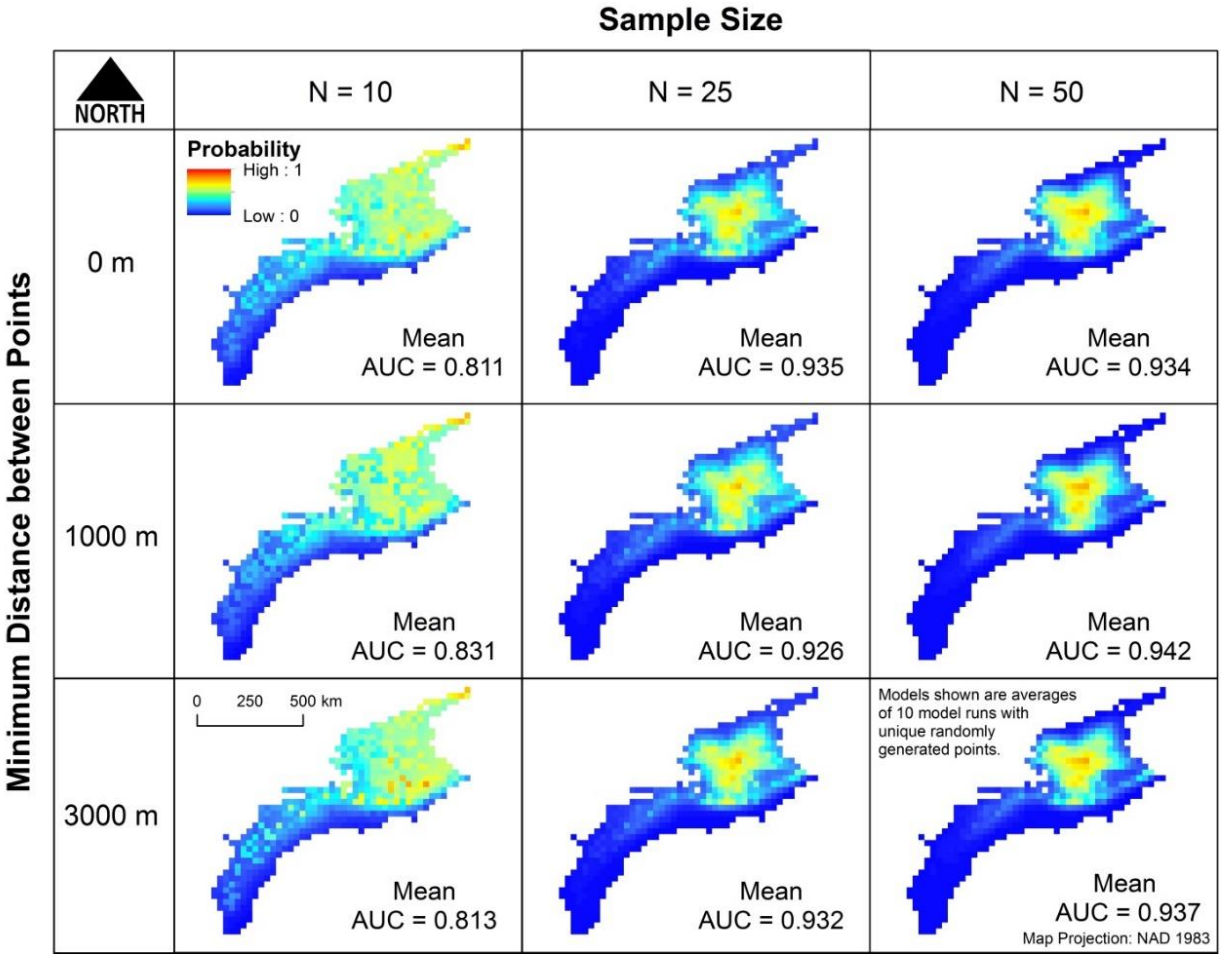


Figure 8. Final averaged maximum entropy models for red hake in spring 2014 using 10 uniquely and randomly generated input point datasets for each model type (combination of sample size and dispersion between points).

Table 4. AUC score descriptive statistics for average red hake fall 2013 Maxent models

N	Dispersion (km)	Mean	Min	Max	STD	SE
10	0	0.8483	0.792	0.917	0.033327	0.010539
	1	0.8391	0.773	0.912	0.046995	0.014861
	3	0.858	0.821	0.908	0.035867	0.011342
25	0	0.9322	0.897	0.962	0.023869	0.007548
	1	0.936	0.887	0.952	0.019442	0.006148
	3	0.9462	0.919	0.958	0.011783	0.003726
50	0	0.945	0.937	0.952	0.004922	0.001556
	1	0.9435	0.936	0.952	0.005104	0.001614
	3	0.944	0.936	0.952	0.005598	0.00177

Table 5. AUC score descriptive statistics for average red hake spring 2014 Maxent models

N	Dispersion (km)	Mean	Min	Max	STD	SE
10	0	0.8114	0.687	0.865	0.059384	0.018779
	1	0.8315	0.776	0.917	0.046933	0.014842
	3	0.8131	0.717	0.886	0.066382	0.020992
25	0	0.9355	0.896	0.959	0.022989	0.00727
	1	0.9261	0.859	0.955	0.032233	0.010193
	3	0.9328	0.858	0.956	0.027912	0.008826
50	0	0.9345	0.927	0.95	0.007276	0.002301
	1	0.9422	0.928	0.956	0.008311	0.002628
	3	0.9372	0.925	0.951	0.008257	0.002611

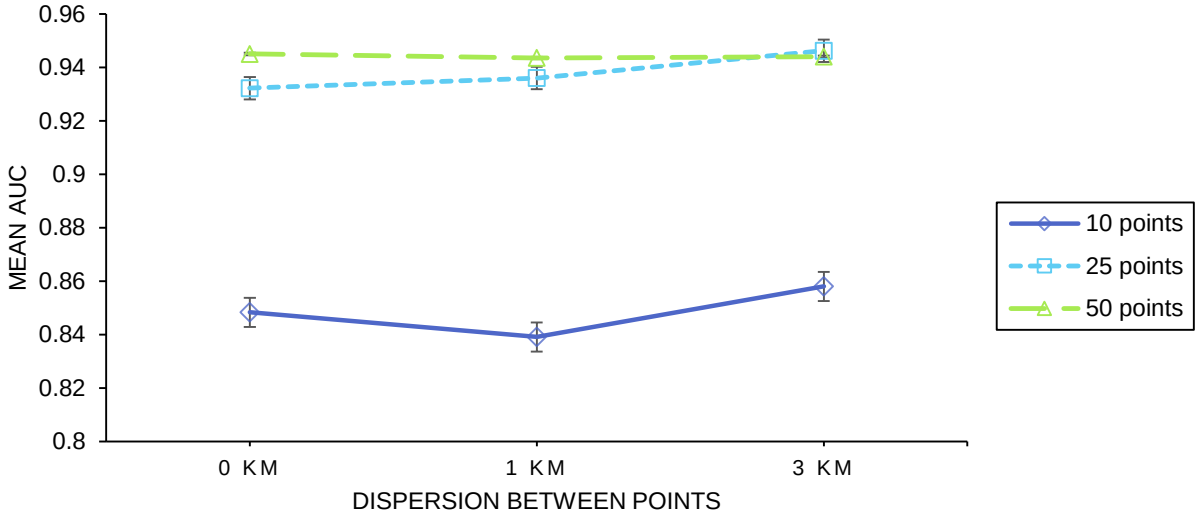


Figure 9. Fall 2013 red hake mean proxy Maxent model AUC scores graphed in terms of point clustering. Error bars represent one standard error of the mean.

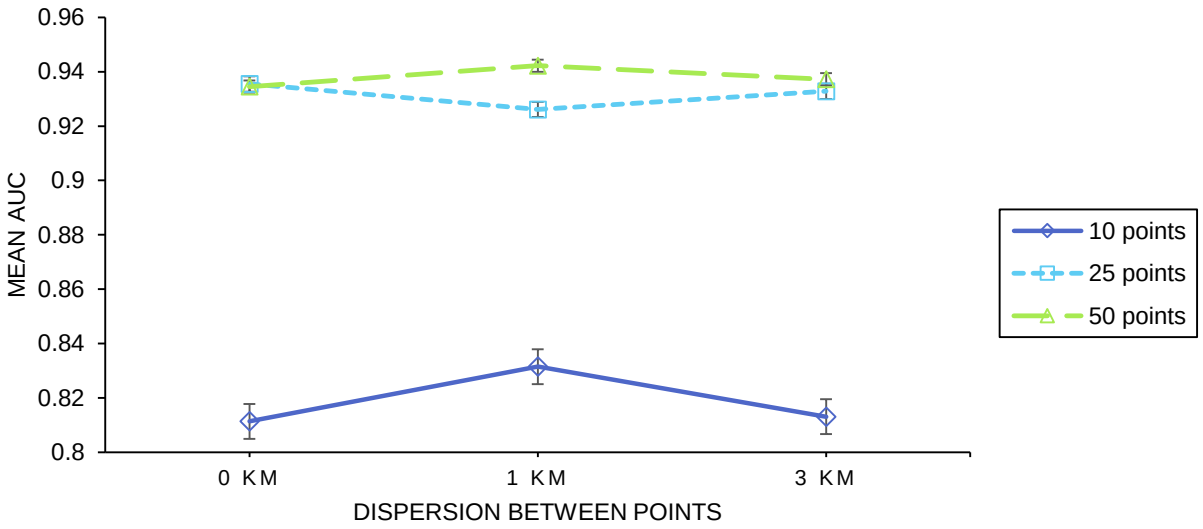


Figure 10**.** Spring 2014 red hake mean proxy Maxent model AUC scores graphed in terms of point clustering. Error bars represent one standard error of the mean.

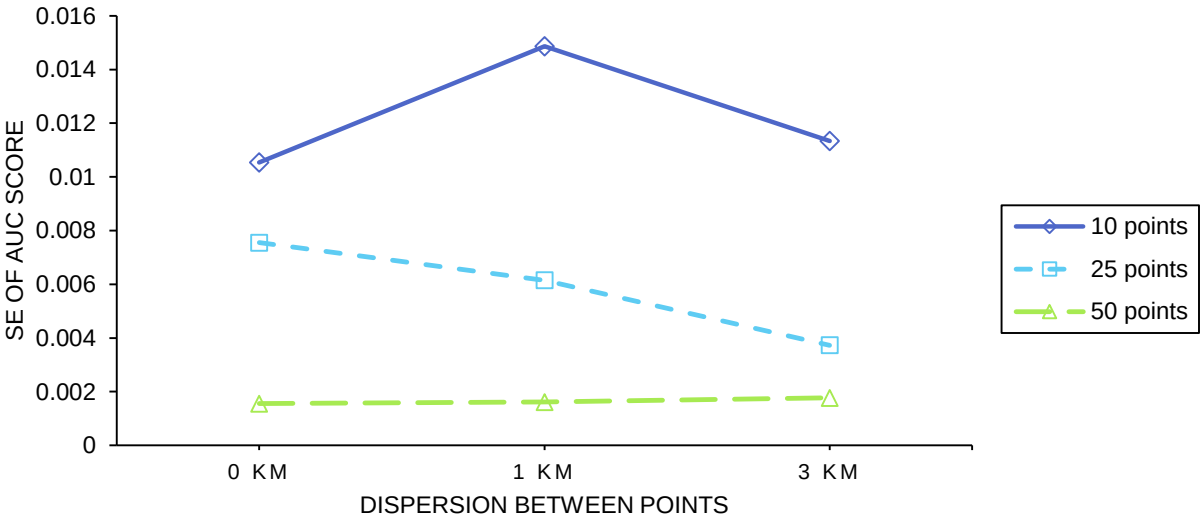


Figure 11. Standard error of fall 2013 red hake mean proxy Maxent model AUC scores graphed in terms of point clustering

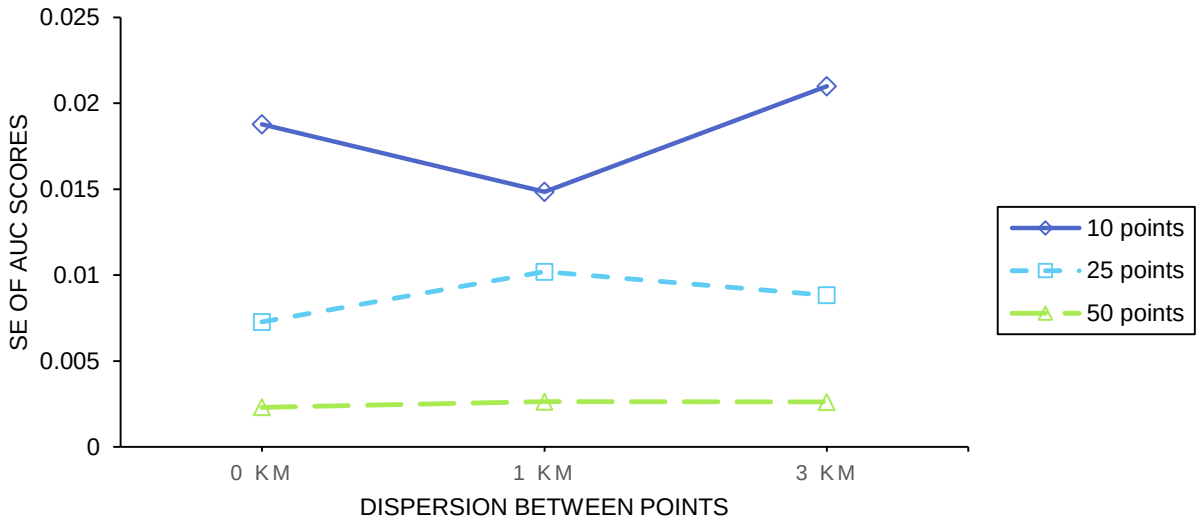


Figure 12. Standard error of spring 2014 red hake mean proxy Maxent model AUC scores graphed in terms of point clustering

3.5. Model comparison

In order to compare the trawl survey Maxent model to the proxy data Maxent models, the individual proxy models were subtracted from the trawl model. Values farther from 0 (darker values in Figures 13 and 14) are indicative of a larger difference between the two models. Positive values mean that the proxy model is underestimating in comparison to the trawl model and negative values mean that the proxy model is overestimating in comparison to the trawl model. For both seasons, red hake proxy data models underestimated the probability of presence off the continental shelf and on the Mid-Atlantic Bight (Figures 13 and 14). The mean values of all difference grids were positive, suggesting that all the Maxent models using proxy data underestimate the area of the red hake ecological niche in comparison to the trawl data presence model. As sample size for proxy data increased, so did the mean value of the difference grid (Figures 15 and 16); this may be due to the increase in model specificity as the sample size

increased. Point clustering appears to have no effect on the difference between models; three levels of clustering is not sufficient to show a trend.

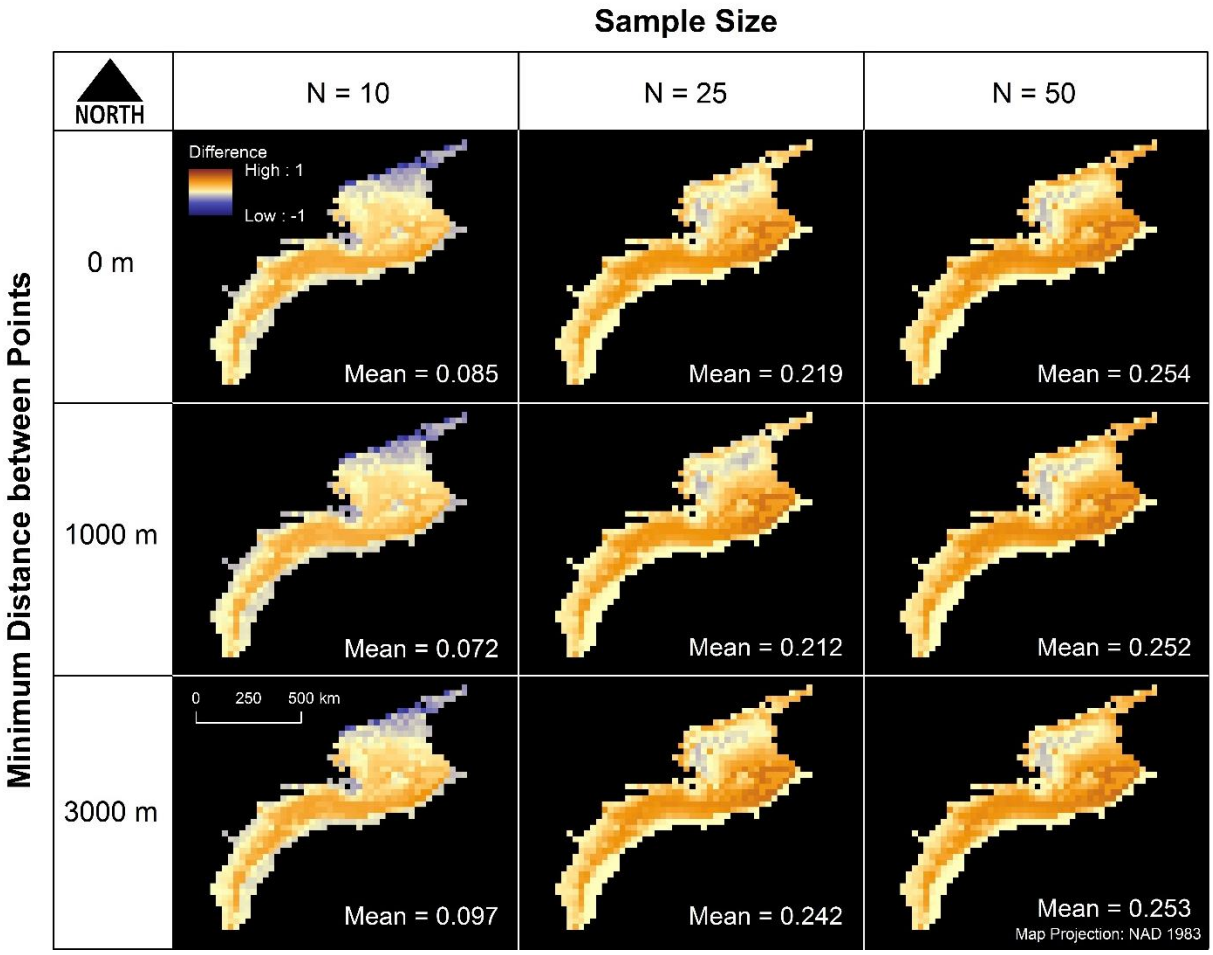


Figure 13. Difference between fall 2013 maximum entropy model and averaged maximum entropy models using proxy data of differing combinations of sample size and clustering. All averaged proxy models were subtracted from the maximum entropy model using presence data from the trawl survey (see Figure 7). Darker colors represent a larger difference between the two models.

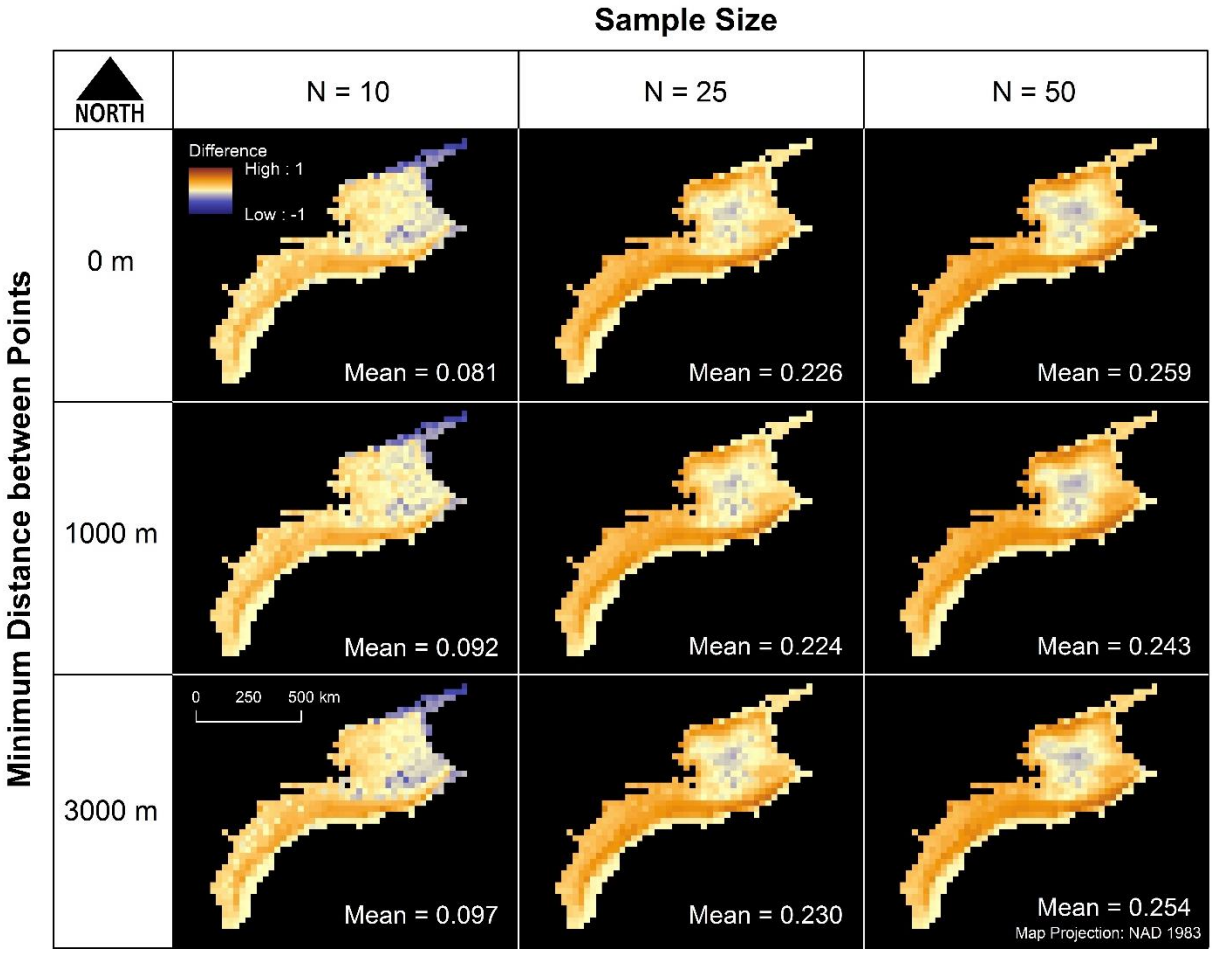


Figure 14. Difference between spring 2014 maximum entropy model and averaged maximum entropy models using proxy data of differing combinations of sample size and clustering. All averaged proxy models were subtracted from the maximum entropy model using presence data from the trawl survey (see Figure 8). Darker colors represent a larger difference between the two models.

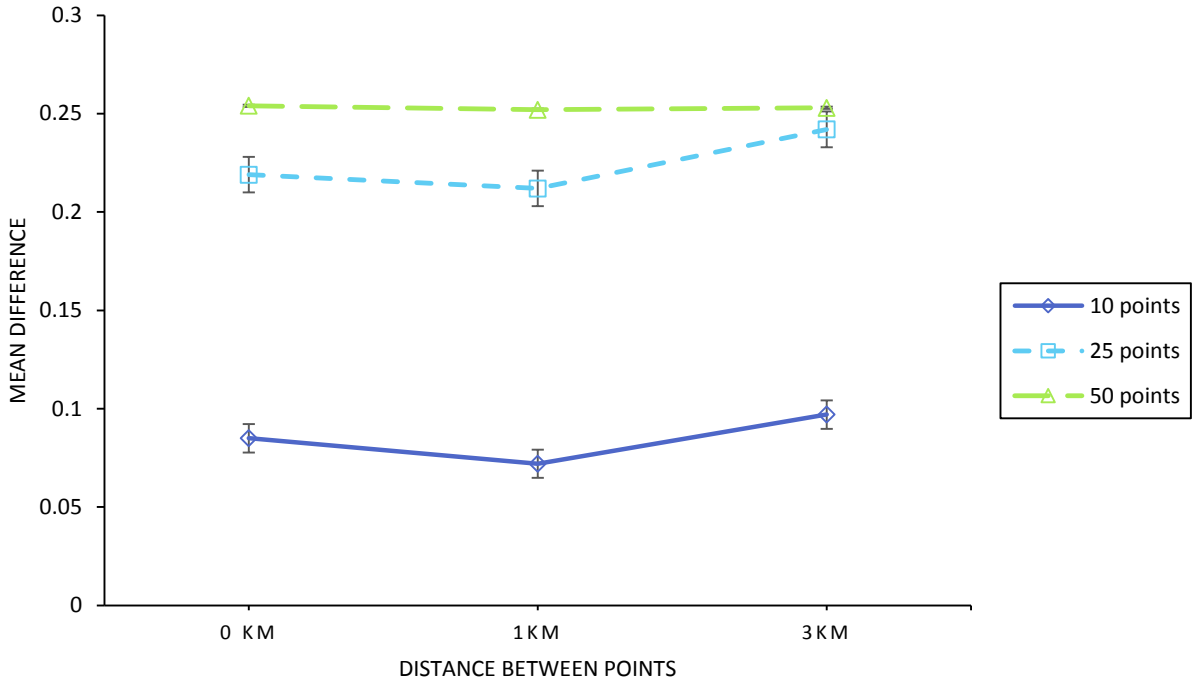


Figure 15. Average value of difference raster from trawl survey Maxent model for fall 2013 red hake proxy models. Error bars represent one standard error of the mean.

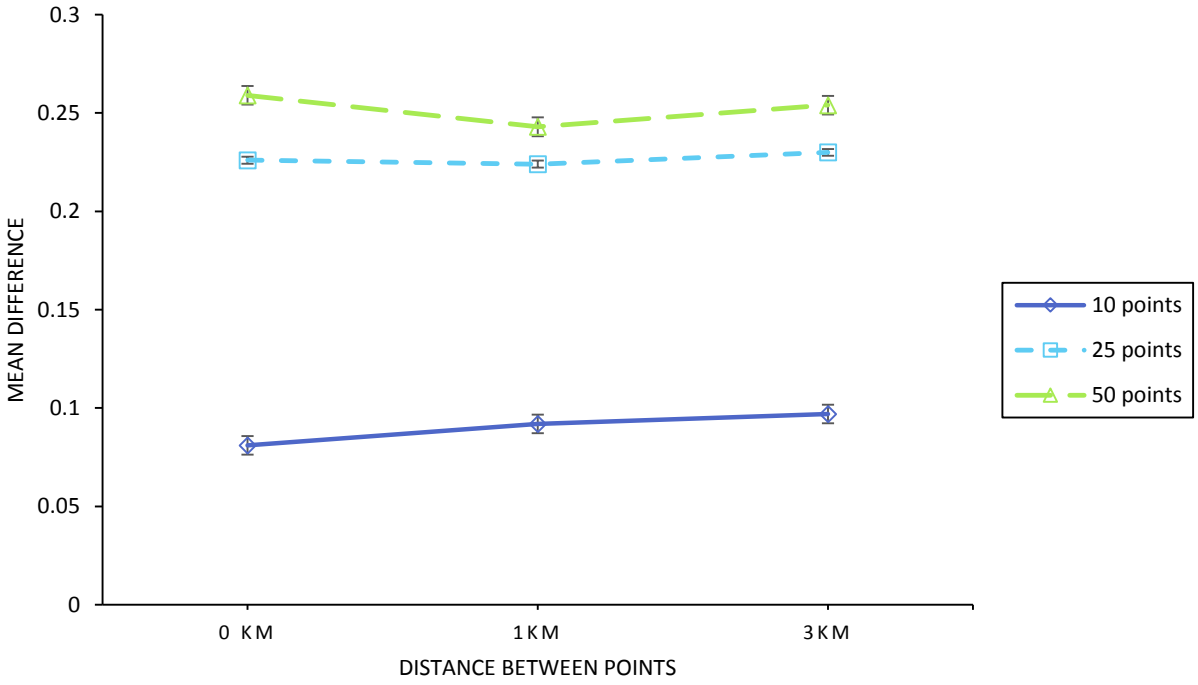


Figure 16. Average value of difference raster from trawl survey Maxent model for spring 2014 red hake proxy models. Error bars represent one standard error of the mean.

4. Discussion

Our results show that data on moving species collected by commercial fishermen in the Gulf of Maine are suitable for species distribution modeling. As previously mentioned, citizen science data have been essential for documenting poleward range shifts for numerous terrestrial taxa across the world (Hickling *et al*., 2006; Parmesan and Yohe, 2003; Walther *et al*., 2002). Ecological niche modeling, or species distribution modeling, is a quantitative way of estimating species geographic ranges from occurrence records and the environmental conditions found there (Elith and Leathwick, 2009). All proxy data models were judged as good (AUC > 0.75). As expected, an increase in sample size led to higher quality models (higher AUC) under all circumstances. Thus, if a phone application is created, the greater amount of input from fishermen, the more robust the model outputs will be.

For many poorly known species, SDM is an essential tool in providing geographic range estimates and determining habitat preferences (Beck *et al.*, 2014). However, it is understood that ecological niche models are extremely sensitive to the distortion of observed environmental conditions in specimen records caused by spatial bias (Dudík and Phillips, 2005; Lintz *et al.*, 2013; Phillips *et al.*, 2009). It is for this reason that we evaluated the potential for using citizen science data from commercial fishermen by assessing multiple sample sizes and spatial distributions of proxy data. Our results showed that spatial bias in specimen distributions did not reduce the quality (AUC) of the predictive distribution models. Nonetheless, some contend that the AUC is flawed as a measure of predictive quality, as it can be influenced by the selected extent, is not consistent with other evaluation standards, and (when applied to presence-only data) does not describe "true" AUC (Barve *et al.*, 2001; Jimenez-Valverde, 2001; Lobo *et al*., 2008; Peterson *et al*., 2008). The AUC statistics of models with differing levels of spatial clustering in the input proxy data did not

exhibit any clear trend, which may be attributed to the fact that the AUC itself may be misled by data bias and therefore unable to expose actual changes in quality (Beck *et al*., 2014).

In terms of predictive quality, the "citizen science" models did fail to identify red hake's preference for deep waters off the continental shelf for both seasons indicated by the initial GAMs used to select areas of habitat. The Maxent model using presence-only data from the trawl survey also failed to identify this preference, indicating that the issue is not due to poor presence data quality, but rather is a function of the lack of absence location data and the maximum entropy model itself.

In general, the trawl survey Maxent model did a poor job of predicting red hake presence probability. One fundamental assumption of Maxent, and other SDMs, is that the entire study area has been systematically or randomly sampled (Phillips *et al*., 2009; Royle *et al*., 2012). In the case of the trawl survey, the entire study region was systematically sampled, via a random stratified sampling strategy. We suspect that the trawl models may have performed poorly because red hake presence points appeared ubiquitously throughout the study region, making identification of the species' ecological niche environmental parameters difficult. As for the proxy data models, systematic or random sampling did not occur, as presence points were limited by where groundfishing took place at the time of trawl sampling. Fishing only occurred in a small region of the study area centered in the Gulf of Maine (refer to Figure 4), which is problematic because Maxent assumes that the entire region has been sampled. Our difference results between the trawl and proxy Maxent models indicate that using concentrated data from fishing areas can lead to underestimation of species presence probability in regions where no fishing occurred. Spatial bias such as using only Gulf of Maine fishing data can lead to environmental bias because of the over-representation of certain environmental features of the comprehensively sampled regions. In turn,

spatial clustering frequently results in autocorrelation of the model residuals and affects model quality by inflating its accuracy (Veloz, 2009). This means that statistical significance may be assigned to environmental predictor variables in the SDMs that are merely representative of the region of intensive survey, which results in auxiliary spatial extrapolation errors (Kramer-Schadt *et al*., 2013).

Consequently, any actual attempt to model citizen science data from commercial fishermen in New England should also include data from Mid-Atlantic fishermen. If the study region were clipped to areas north of Cape Cod, model performance and accuracy would improve, but the models would fail to fully demonstrate any distributional shifts poleward from the Mid-Atlantic Bight, which is the purpose for collecting data from commercial fishermen in the first place. Collecting data from fishermen in both regions would be optimal and allow for complete modeling of the study region; it may also improve overall model quality and the legitimacy of any findings regarding distributional changes.

Another option would be to combine the citizen science dataset with the trawl survey data. Scientists are just now beginning to see the benefits of combining data from separate, independent sources (Link *et al*., 2008). In fact, certain collaborative efforts are developing that seek to combine and integrate data for use in an assortment of analyses (Kelling *et al*., 2009). It is vital to acknowledge that scientific and conservation goals are best served by improving upon existing projects that concern datasets that may be merged (Dickinson *et al*., 2010). Combining datasets may produce models more accurate than either the trawl survey or citizen science datasets could alone. If this strategy is selected, modelers must be very careful in how they go about selecting which data should be included in their models in order to address issues of sampling bias. Dudík and Phillips (2005) and Phillips *et al*. (2009) have both suggested using a bias file that assigns the

probability of background environmental samples to regions that have truly been well sampled. This kind of file can be created using a priori knowledge of sampling intensity from features of population density (Ballesteros-Mejia *et al*., 2013), or in this case from fishing density (from VTR data or Vessel Monitoring System/VMS outputs). Thus, combining datasets and implementing a bias file may also be a feasible option for improving upon the SDMs produced from NEFSC trawl survey data alone.

5. Conclusion

Data collected by New England *are* suitable for maximum entropy species distribution modeling and should be collected by the GMRI and the Island Institute. Since many of the moving species' extents stretch into the Mid-Atlantic, fishermen from this region should also be encouraged to submit data through the proposed phone application. Presence-only data from fishermen will likely demonstrate different trends in each species' movement from models using data from the NEFSC trawl survey. Since fishermen feel that the NEFSC survey is failing to accurately depict these changes in distribution, it is possible that models from citizen science data will provide new insight, especially for seasons during which no trawl survey occurs.

References

Ames, E. P. (2004). Atlantic cod stock structure in the Gulf of Maine. *Fisheries, 29*(1), 10-28.

Azarovitz, T. R. (1981). A brief historical review of the Woods Hole Laboratory trawl survey time series. *Canadian Special Publication of Fisheries and Aquatic Sciences, 58*, 62-67.

Ballesteros-Mejia, L., Kitching, I. J., Jetz, W., Nagel, P., & Beck, J. (2013). Mapping the biodiversity of tropical insects: species richness and inventory completeness of African sphingid moths. *Global Ecology and Biogeography, 22*(5), 586-595.

Barve, N., Barve, V., Jiménez-Valverde, A., Lira-Noriega, A., Maher, S. P., Peterson, A. T., . . . Villalobos, F. (2011). The crucial role of the accessible area in ecological niche modeling and species distribution modeling. *Ecological Modelling, 222*(11), 1810-1819. doi: http://dx.doi.org/10.1016/j.ecolmodel.2011.02.011

Baum, J. K., & Worm, B. (2009). Cascading top-down effects of changing oceanic predator abundances. *Journal of Animal Ecology, 78*, 699-714.

Beck, J., Böller, M., Erhardt, A., & Schwanghart, W. (2014). Spatial bias in the GBIF database and its effect on modeling species' geographic distributions. *Ecological Informatics, 19*, 10-15.

Bell, R. J., Richardson, D. E., Hare, J. A., Lynch, P. D., & Fratantoni, P. S. (2014). Disentangling the effects of climate, abundance, and size on the distribution of marine fish: an example based on four stocks from the Northeast US shelf. *ICES Journal of Marine Science: Journal du Conseil*, fsu217.

Beyer, H. L. (2012). Geospatial Modelling Environment (Version 0.7.3.0). (software).

de Solla, S., Shirose, L., Fernie, K., Barrett, G., Brousseau, C., & Bishop, C. (2005). Effect of sampling effort and species detectability on volunteer based anuran monitoring programs. *Biological Conservation, 121*, 585-594.

Dobbs, D. (2000). *The Great Gulf: Fishermen, Scientists, and the Struggle to Revive the World's Greatest Fishery*. Washington, D.C: Island Press.

Dudík, M., Phillips, S. J., & Schapire, R. E. (2005). *Correcting sample selection bias in maximum entropy density estimation.* Paper presented at the Advances in neural information processing systems.

Elith, J., H. Graham, C., P. Anderson, R., Dudík, M., Ferrier, S., Guisan, A., . . . E. Zimmermann, N. (2006). Novel methods improve prediction of species' distributions from occurrence data. *Ecography, 29*(2), 129-151. doi: 10.1111/j.2006.0906-7590.04596.x

Elith, J., & Leathwick, J. R. (2009). Species distribution models: ecological explanation and prediction across space and time. *Annual review of ecology, evolution, and systematics, 40*(1), 677.

ESRI (2013). ArcGIS Desktop: Release 10.2. Redlands, CA: Environmental Systems Research Institute.

Fielding, A. H., & Bell, J. F. (1997). A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental conservation, 24*(01), 38-49.

Fitzpatrick, M. C., Gotelli, N. J., & Ellison, A. M. (2013). MaxEnt versus MaxLike: empirical comparisons with ant species distributions. *Ecosphere, 4*(5), art55. doi: 10.1890/ES13-00066.1

Fogarty, M., Incze, L., Hayhoe, K., Mountain, D., & Manning, J. (2008). Potential climate change impacts on Atlantic cod (Gadus morhua) off the northeastern USA. *Mitigation and Adaptation Strategies for Global Change, 13*(5-6), 453-466. doi: http://dx.doi.org/10.1007/s11027-007-9131-4

Genet, K., & Sargent, L. (2003). Evaluation of methods and data quality from a volunteer-based amphibian call survey. *Wildlife Society Bulletin, 31*, 703-714.

Greene, K. (2002). Bigger populations needed for sustainable harvests. *Science, 296*, 1229-1230.

Gutowsky, L. F., Gobin, G. J., Burnett, N. J., Chapman, J. M., Stoot, L. J., & Bliss, S. (2013). Smartphones and Digital Tablets: Emerging Toots for Data Collection and Education in Fisheries. *Fisheries, 38*(10), 455-461.

Hare, J. A., Wuenschel, M. J., & Kimball, M. E. (2012). Projecting Range Limits with Coupled Thermal Tolerance - Climate Change Models: An Example Based on Gray Snapper (<italic>Lutjanus griseus</italic>) along the U.S. East Coast. *PloS one, 7*(12), e52294. doi: 10.1371/journal.pone.0052294

Hickling, R., Roy, D., Hill, J., Fox, R., & Thomas, C. (2006). The distributions of a wide range of taxonomic groups are expanding polewards. *Conservation Biology, 21*, 534-539.

Hudson, M., & Peros, J. (2013). Preparing for Emerging Fisheries: An Overview of Mid-Atlantic Stocks on the Move: Gulf of Maine Research Institute.

Institute, I. (2008). A Climate of Change: A Preliminary Assessment of Fishermen's Observations on a Dynamic Fishery: Island Institute.

Ji, R., Davis, C. S., Chen, C., Townsend, D. W., Mountain, D. G., & Beardsley, R. C. (2007). Influence of ocean freshening on shelf phytoplankton dynamics. *Geophysical Research Letters, 34*(24), L24607. doi: 10.1029/2007GL032010

Jiménez-Valverde, A., Acevedo, P., Barbosa, A. M., Lobo, J. M., & Real, R. (2013). Discrimination capacity in species distribution models depends on the representativeness of the environmental domain. *Global Ecology and Biogeography, 22*(4), 508-516. doi: 10.1111/geb.12007

Jones, M. C., Dye, S. R., Pinnegar, J. K., Warren, R., & Cheung, W. W. (2012). Modelling commercial fish distributions: Prediction and assessment using different approaches. *Ecological Modelling, 225*, 133-145.

Kelling, S., Hochachka, W. M., Fink, D., Riedewald, M., Caruana, R., & al, e. (2009). Data-intensive science: a new paradigm of biodiversity studies *BioScience, 59*.

Kramer-Schadt, S., Niedballa, J., Pilgrim, J. D., Schröder, B., Lindenborn, J., Reinfelder, V., . . . Augeri, D. M. (2013). The importance of correcting for sampling bias in MaxEnt species distribution models. *Diversity and Distributions, 19*(11), 1366-1379.

Link, W. A., R, S. J., & Niven, D. K. (2008). Combining breeding bird survey and Christmas Bird Count data to evaluate seasonal components of population change in norther bobwhite. *Journal of Wildlife Management, 72*, 44-51.

Lintz, H. E., Gray, A. N., & McCune, B. (2013). Effect of inventory method on niche models: Random versus systematic error. *Ecological Informatics, 18*, 20-34.

Lobo, J. M., Jiménez-Valverde, A., & Real, R. (2008). AUC: a misleading measure of the performance of predictive distribution models. *Global Ecology and Biogeography, 17*(2), 145-151.

Lotz, A., & Allen, C. (2007). Observer bias in anuran call surveys. *Journal of Wildlife Management, 71*, 675-679.

MERCINA. (2013). Remote climate forcing of decadal-scale regime shifts in Northwest Atlantic shelf ecosystems. *Limnology and Oceanography, 58*, 803-816.

MGET. (2014). Marine Geospatial Ecology Tools <http://mgel.env.duke.edu/mget>.

Miller-Rushing, A., Primack, R., & Bonney, R. (2012). The history of public participation in ecological research. *Frontiers in Ecology and the Environment, 10*(6), 285-290.

Mills, K. E., Pershing, A. J., Brown, C. J., Chen, Y., Chiang, F., Holland, D. S., . . . Wahle, R. A. (2013). Fisheries management in a changing climate: lessons from the 2012 ocean heat wave. *Oceanography, 26*(2), 191-195.

Mills, K. E., Pershing, A. J., Brown, C. J., Chen, Y., Chiang, F.-S., Holland, D. S., . . . Thomas, A. C. (2013). Fisheries Management in a Changing Climate Lessons from the 2012 Ocean Heat Wave in the Northwest Atlantic. *Oceanography, 26*(2), 191-195.

Mountain, D. G., & Kane, J. (2010). Major changes in the Georges Bank ecosystem, 1980s to the 1990s. *Marine Ecology Progress Series, 398*, 81-91. doi: 10.3354/meps08323

NODC. (2014). <http://www.nodc.noaa.gov/sog/ghrsst/accessdata.html>.

NROC. (2014). Northeast Ocean Data Portal <http://www.northeastoceandata.org/data/data-download/>.

Nye, J. A., Link, J. S., Hare, J. A., & Overholtz, W. J. (2009). Changing spatial distribution of fish stocks in relation to climate and population size on the Northeast United States continental shelf. *Marine Ecology Progress Series, 393*, 111-129.

Parmesan, C., & Yohe, G. (2003). A globally coherent fingerprint of climate change impacts across natural systems. *Nature, 421*, 37-42.

Pearson, R. G., Raxworthy, C. J., Nakamura, M., & Townsend Peterson, A. (2007). Predicting species distributions from small numbers of occurrence records: a test case using cryptic geckos in Madagascar. *Journal of Biogeography, 34*(1), 102-117.

Perkins, N. J., & Schisterman, E. F. (2006). The Inconsistency of "Optimal" Cutpoints Obtained using Two Criteria based on the Receiver Operating Characteristic Curve. *American Journal of Epidemiology, 163*(7), 670-675. doi: 10.1093/aje/kwj063

Pershing, A. J., Greene, C. H., Jossi, J. W., O'Brien, L., Brodziak, J. K. T., & Bailey, B. A. (2005). Interdecadal variability in the Gulf of Maine zooplankton community with potential impacts on fish recruitment. *ICES Journal of Marine Science, 62*(1511-1523).

Peterson, A. T., Papeş, M., & Soberón, J. (2008). Rethinking receiver operating characteristic analysis applications in ecological niche modeling. *Ecological Modelling, 213*(1), 63-72.

Phillips, S. J., Anderson, R. P., & Schapire, R. E. (2006). Maximum entropy modeling of species geographic distributions. *Ecological Modelling, 190*(3–4), 231-259. doi: http://dx.doi.org/10.1016/j.ecolmodel.2005.03.026

Phillips, S. J., Dudík, M., Elith, J., Graham, C. H., Lehmann, A., Leathwick, J., & Simon, F. (2009). Sample Selection Bias and Presence-Only Distribution Models: Implications for Background and Pseudo-Absence Data. *Ecological Applications, 19*(1), 181-197. doi: 10.2307/27645958

Phillips, S. J., Dudík, M., & Schapire, R. E. (2004). *A maximum entropy approach to species distribution modeling.* Paper presented at the Proceedings of the twenty-first international conference on Machine learning.

Pierce, B., & Gutzwiller, K. (2007). Interobserver variation in frog call surveys. *Journal of Herpetology, 41*, 424-429.

Pinsky, M., & Fogarty, M. (2012). Lagged social-ecological responses to climate and range shifts in fisheries. *Climate Change, 115*, 883-891.

Royle, J. A., Chandler, R. B., Yackulic, C., & Nichols, J. D. (2012). Liklihood analysis of species occurrence probability from presence-only data for modelling species distributions. *Methods in Ecology and Evolution, 3*, 545-554.

Sing, T., Sander, O., Beerenwinkel, N., & Lengauer, T. (2007). <http://rocr.bioinf.mpi-sb.mpg.de/ROCR>.

Singer, L., Arnold, S., Battista, N., & Deese, H. (2013). A Climate of Change - Climate Change and New England Fisheries: Observations, Impacts, and Adaptation Strategies: The Island Institute.

Valavanis, V. D., Pierce, P., Zuur, A., Palialexis, A., Saveliev, A., Katara, I., & Wang, J. (2008). Modeling of essential fish habitat based on remote sensing, spatial analysis, and GIS. *Developments in Hydrobiology, 203*, 5-20.

Veloz, S. D. (2009). Spatially autocorrelated sampling falsely inflates measures of accuracy for presence-only niche models. *Journal of Biogeography, 36*, 2290-2229.

Walther, G., Post, E., Convey, P., Menzel, A., & Parmesan, C. (2002). Ecological responses to recent climate change. *Nature, 416*, 389-395.

Wier, L., Royle, J. A., Nanjappa, P., & Jung, R. (2005). Modeling anuran detection and site occupancy on North American Amphibian Monitoring Program (NAAMP) routes in Maryland. *Journal of Herpetology, 39*, 627-639.

Acknowledgements

This research would not have been possible without the guidance and support of my advisors Dr. Patrick Halpin and Dr. Kathy Mills. A sincere thank you to Dr. Andy Pershing and Jonathan Peros of the GMRI for assistance in developing the project in its initial stages. Thank you to Kelley McGrath of NOAA for processing my VTR data requests. Special thanks to John Fay for teaching me how to code in Python and ArcPy, which was essential to completing this work. Finally, thank you to my fiancé Stuart Sheehan for his unyielding support and encouragement.

Appendix

Item 1. Python script created to run general additive models in an ArcGIS tool in a toolbox

```
##---------------------------------------------------------------------------------
## Script Name:  Data_Extraction.py
##
## Description:  New England GAM tool for Scup, Black Sea Bass, and Red Hake.
##               This tool will extract presence and absence points for a user
##               selected species and season and create a shapefile of those points.
##               Then it will sample input rasters (environmental data) for the points
##               in the shapefile and and produce a DBase table of the values. Next
##               it will remove any points for which any environmental data is
##               missing.
##               This table will serve as the input for the Fit GAM tool in the Marine
##               Geospatial Ecology Toolbox.
##
##               Note: The user must have the MGET toolbox installed on his or her
##               computer for this tool to function. See http://mgel.env.duke.edu/mget
##               for more information.
##
##               This tool also requires the spatial analyst extension in ArcGIS.
##
## Created:      November 2014
## Author:       Julia Livermore - julia.livermore@duke.edu (for Master's Research)
##---------------------------------------------------------------------------------

# Import system modules
import arcpy, os, sys
from arcpy import env
from arcpy.sa import *

# Set environmental settings
env.workspace = sys.argv[1]
env.scratchWorkspace = sys.argv[2]
env.overwriteOutput = True

# Check out the ArcGIS Spatial Analyst extension license
arcpy.CheckOutExtension("Spatial")

#---------------------------------------------------------------------------
# Data Extraction from NOAA Trawl Survey Data
arcpy.AddMessage("Extracting data.")

#Get user input on species and season (only options are Spring 2014 and Fall 2103)
## These options will be explicitly described in the ArcGIS tool script.
Species = arcpy.GetParameterAsText(2)
    ### BSB, RED HAKE, and SCUP
Season = arcpy.GetParameterAsText(3)
    ### SPRING, FALL
    ### Explain that spring is spring 2014 and fall is fall 2013

# Create string for simpler paths
SpeciesSeason = str(Species + "_" + Season)

# Create a shapefile of absence points from the survey data
```

```
# Process: Create Feature Class
arcpy.CreateFeatureclass_management(env.scratchWorkspace, SpeciesSeason + "_ABS.shp",
"POINT", env.workspace + "\\Pres_Abs_Template.shp",
                                    "DISABLED", "DISABLED",env.workspace +
"\\Spatial_Reference.prj", "", "0", "0", "0")

# Fields OBJECTID, SEASON, SPECIES, PRES_ABS, BEGLON, and BEGLAT added to feature
# class from the template.
absFC = env.scratchWorkspace + "\\" + SpeciesSeason + "_ABS.shp"

# Create an input cursor for the feature class so that we can add feature records
cur = arcpy.InsertCursor(absFC)

# Set input file to read the data from based on user input parameters
inputFile = env.workspace + "\\Trawl_Data.csv"

# Extract entries from folder into a list based on user inputs
## Open csv file for reading
inputFileObj = open(inputFile,'r')
## Start with first line and begin while loop through document.
lineString = inputFileObj.readline()
while lineString:
    # Only transfer data from lines including the user-selected species and season
    if ((Species in lineString) and (Season in lineString)):
        # Parse line into a list
        lineData = lineString.split(',')
        if (lineData[3] is "0"):
            # Extract attributes from the datum header line
            objectID = lineData[0]
            obsSpecies = lineData[2]
            obsSeason = lineData[8]
            presAbs = lineData[3]
            estYear = lineData[4]
            begLong = lineData[20]
            begLat = lineData[18]

            try:
                # Create a point object from the new feature class
                obsPoint = arcpy.Point()
                obsPoint.X = begLong
                obsPoint.Y = begLat

                # Create a feature object to add to the feature class
                featObj = cur.newRow()

                # Set the feature's shape and other attribute values
                featObj.shape = obsPoint
                featObj.setValue("OBJECTID",objectID)
                featObj.setValue("SPECIES",obsSpecies)
                featObj.setValue("PRES_ABS",presAbs)
                featObj.setValue("EST_YEAR",estYear)
                featObj.setValue("SEASON",obsSeason)
                featObj.setValue("BEGLON",begLong)
                featObj.setValue("BEGLAT",begLat)

                # Commit the feature to the feature class
                cur.insertRow(featObj)
            except Exception as e:
                print e, "Error adding point" + objectID + "to the file."

    # Move to the next line to continute the while loop.
    lineString = inputFileObj.readline()
```

```
# Close the file object and delete cursor
inputFileObj.close()
del cur

# Create a shapefile of absence points from the survey data
# Set Local variables:
Pres_Abs_Template_shp = env.workspace + "\\Pres_Abs_Template.shp"
outputShapefile = SpeciesSeason + "_PRES.shp"

# Process: Create Feature Class
arcpy.CreateFeatureclass_management(env.scratchWorkspace, outputShapefile, "POINT",
Pres_Abs_Template_shp, "DISABLED", "DISABLED",
                                    env.workspace + "\\Spatial_Reference.prj","", "0",
"0", "0")

# Fields OBJECTID, SEASON, SPECIES, PRES_ABS, BEGLON, and BEGLAT added to feature
# class from the template.
presFC = env.scratchWorkspace + "\\" + SpeciesSeason + "_PRES.shp"

# Create an input cursor for the feature class so that we can add feature records
cur = arcpy.InsertCursor(presFC)

# Set input file to read the data from based on user input parameters
inputFile = env.workspace + "\\Trawl_Data.csv"

# Extract entries from folder into a list based on user inputs
## Open csv file for reading
inputFileObj = open(inputFile,'r')
## Start with first line and begin while loop through document.
lineString = inputFileObj.readline()
while lineString:
    # Only transfer data from lines including the user-selected species and season
    if ((Species in lineString) and (Season in lineString)):
        # Parse line into a list
        lineData = lineString.split(',')
        if (lineData[3] is "1"):
            # Extract attributes from the datum header line
            objectID = lineData[0]
            obsSpecies = lineData[2]
            obsSeason = lineData[8]
            presAbs = lineData[3]
            estYear = lineData[4]
            begLong = lineData[20]
            begLat = lineData[18]

            try:
                # Create a point object from the new feature class
                obsPoint = arcpy.Point()
                obsPoint.X = begLong
                obsPoint.Y = begLat

                # Create a feature object to add to the feature class
                featObj = cur.newRow()

                # Set the feature's shape and other attribute values
                featObj.shape = obsPoint
                featObj.setValue("OBJECTID",objectID)
                featObj.setValue("SPECIES",obsSpecies)
                featObj.setValue("PRES_ABS",presAbs)
                featObj.setValue("EST_YEAR",estYear)
                featObj.setValue("SEASON",obsSeason)
                featObj.setValue("BEGLON",begLong)
                featObj.setValue("BEGLAT",begLat)
```

```
                # Commit the feature to the feature class
                cur.insertRow(featObj)
            except Exception as e:
                print e, "Error adding point" + objectID + "to the file."

    # Move to the next line to continute the while loop.
    lineString = inputFileObj.readline()

# Close the file object and delete cursor
inputFileObj.close()
del cur

arcpy.AddMessage("2 new feature classes have been created in the scratch folder.")

#-------------------------------------------------------------------------------
# Sampling environmental data with datapoints from trawl survey

# Set local variables
sampleMethod = "NEAREST"

if Season is "FALL":
    inRasters = ["bathymetry.img",
                 "bathy_relief.img",
                 "sediments.img",
                 "dist_to_shore.img",
                 "Fall_2013_SST.img"]

else:
    inRasters = ["bathymetry.img",
                 "bathy_relief.img",
                 "sediments.img",
                 "dist_to_shore.img",
                 "Spring_2014_SST.img"]

# Execute Sample
Sample(inRasters, absFC, env.scratchWorkspace + "\\" + SpeciesSeason + "_SampAb.dbf",
sampleMethod)
Sample(inRasters, presFC, env.scratchWorkspace + "\\" + SpeciesSeason + "_SampPr.dbf",
sampleMethod)

arcpy.AddMessage("2 new dBase tables have been created in the scratch folder.")

# Add field for presence-absence values to the tables
arcpy.AddField_management(env.scratchWorkspace + "\\" + SpeciesSeason +
"_SampAb.dbf","PRES_ABS","SHORT")
arcpy.AddField_management(env.scratchWorkspace + "\\" + SpeciesSeason +
"_SampPr.dbf","PRES_ABS","SHORT")

# Fill in values
arcpy.CalculateField_management(env.scratchWorkspace + "\\" + SpeciesSeason +
"_SampAb.dbf","PRES_ABS",0)
arcpy.CalculateField_management(env.scratchWorkspace + "\\" + SpeciesSeason +
"_SampPr.dbf","PRES_ABS",1)

# Merge the two tables into one
arcpy.Merge_management([env.scratchWorkspace + "\\" + SpeciesSeason + "_SampAb.dbf",
                        env.scratchWorkspace + "\\" + SpeciesSeason + "_SampPr.dbf"],
                       env.scratchWorkspace + "\\" + SpeciesSeason + ".dbf")

arcpy.AddMessage("1 new dBase table has been created in the scratch folder.")
#-------------------------------------------------------------------------------
# Select only values where sample data exists for all sampled rasters
```

```
# Set input variables
in_feature = env.scratchWorkspace + "\\" + SpeciesSeason + ".dbf"
out_table = env.scratchWorkspace + "\\" + SpeciesSeason + "no0s.dbf"

if Season is "FALL":
    where_clause = """"bathymetry" < 0 AND "bathy_reli" > 0 AND "sediments" > 0 AND
"dist_to_sh" > 0 AND "Fall_2013_" > 0"""
else:
    where_clause = """"bathymetry" < 0 AND "bathy_reli" > 0 AND "sediments" > 0 AND
"dist_to_sh" > 0 AND "Spring_201" > 0"""

# Execute table select
arcpy.TableSelect_analysis(in_feature, out_table, where_clause)

# Delete all temporary files
arcpy.AddWarning("Deleting temporary files.")

os.remove(env.scratchWorkspace + "\\" + SpeciesSeason + ".dbf")
os.remove(env.scratchWorkspace + "\\" + SpeciesSeason + "_PRES.shp")
os.remove(env.scratchWorkspace + "\\" + SpeciesSeason + "_ABS.shp")
os.remove(env.scratchWorkspace + "\\" + SpeciesSeason + "_PRES.dbf")
os.remove(env.scratchWorkspace + "\\" + SpeciesSeason + "_ABS.dbf")
os.remove(env.scratchWorkspace + "\\" + SpeciesSeason + "_SampPr.dbf")
os.remove(env.scratchWorkspace + "\\" + SpeciesSeason + "_SampAb.dbf")
os.remove(env.scratchWorkspace + "\\" + SpeciesSeason + "_ABS.prj")
os.remove(env.scratchWorkspace + "\\" + SpeciesSeason + "_PRES.prj")

arcpy.AddMessage("One .dbf file has been added to the scratch folder.")
arcpy.AddMessage("The final .dbf table should be used as the input for the Fit GAM
tool in MGET.")
```

Item 2. Python script created by ESRI to select action path – used in ArcGIS tool created by Julia Livermore

```
#*********************************************************************
# Description:
# Tests if a field exists and outputs two booleans:
#   Exists - true if the field exists, false if it doesn't exist
#   Not_Exists - true if the field doesn't exist, false if it does exist
#                (the logical NOT of the first output).
#
# Arguments:
#  0 - Table name
#  1 - Field name
#  2 - Exists (boolean - see above)
#  3 - Not_Exists (boolean - see above)
#
# Created by: ESRI
#*********************************************************************

# Standard error handling - put everything in a try/except block
#
try:

    # Import system modules
    import sys, string, os, arcgisscripting
```

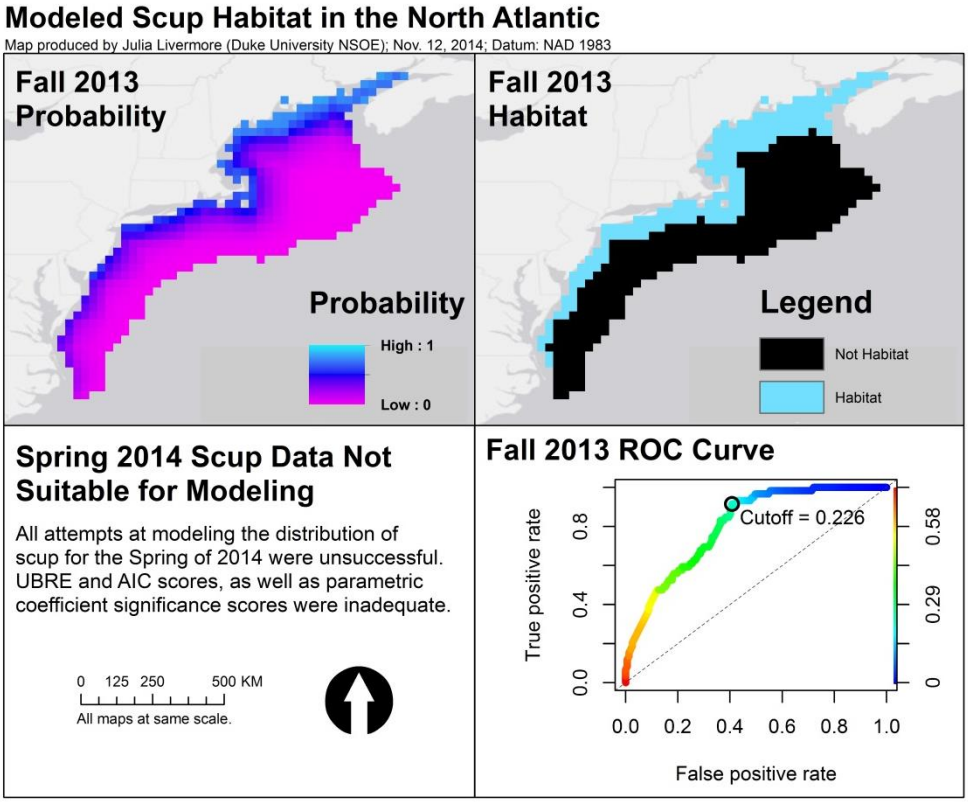
```
    # Create the Geoprocessor object
    gp = arcgisscripting.create()

    # Get input arguments - table name, field name
    #
    in_Table = gp.GetParameterAsText(0)
    in_Field = gp.GetParameterAsText(1)

    # First check that the table exists
    #
    if not gp.Exists(in_Table):
        raise Exception, "Input table does not exist"

    # Use the ListFields function to return a list of fields that matches
    #  the name of in_Field. This is a wildcard match. Since in_Field is an
    #  exact string (no wildcards like "*"), only one field should be returned,
    #  exactly matching the input field name.
    #
    fields = gp.ListFields(in_Table, in_Field)

    # If ListFields returned anything, the Next operator will fetch the
    #  field. We can use this as a Boolean condition.
    #
    field_found = fields.Next()

    # Branch depending on whether field found or not. Issue a
    #  message, and then set our two output variables accordingly
    #
    if field_found:
        gp.AddMessage("Field %s found in %s" % (in_Field, in_Table))
        gp.SetParameterAsText(2, "True")
        gp.SetParameterAsText(3, "False")
    else:
        gp.AddMessage("Field %s not found in %s" % (in_Field, in_Table))
        gp.SetParameterAsText(2, "False")
        gp.SetParameterAsText(3, "True")


# Handle script errors
#
except Exception, errMsg:

    # If we have messages of severity error (2), we assume a GP tool raised it,
    #  so we'll output that.  Otherwise, we assume we raised the error and the
    #  information is in errMsg.
    #
    if gp.GetMessages(2):
        gp.AddError(GP.GetMessages(2))
    else:
        gp.AddError(str(errMsg))
```

Item 3. Python script used in ArcGIS tool to complete general additive model analysis by mapping the areas of habitat determined by the Youden Index

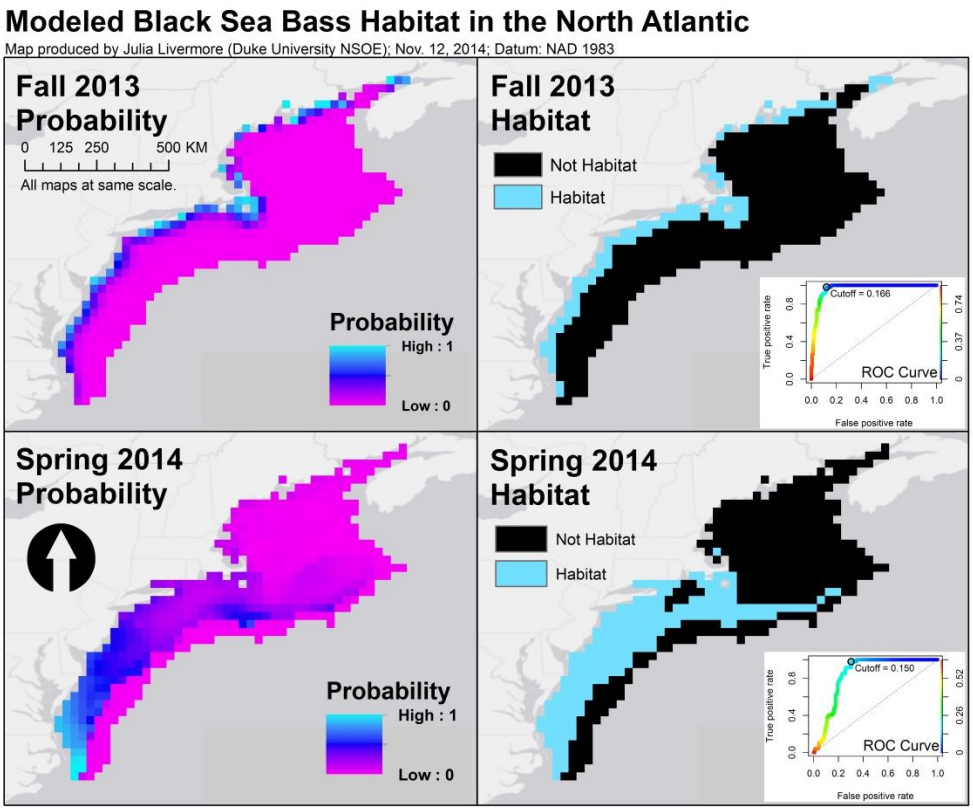
```
##---------------------------------------------------------------------------------
-
## Script Name:  GAM_Raster_Creation.py
##
## Description:  This tool will create a probability raster of the likelihood of
##               encountering the species at each location. A raster of habitat
##               will also be created based on the ROC-determined probability
##               cutoff.
##
##               This tool also requires the spatial analyst extension in ArcGIS.
##
## Created:      November 2014
## Author:       Julia Livermore - julia.livermore@duke.edu (for Master's Research)
##---------------------------------------------------------------------------------
-

# Import system modules
import arcpy, os, sys
from arcpy import env
from arcpy.sa import *

# Set environmental settings
env.workspace = sys.argv[1] ## Set to Data folder again
env.overwriteOutput = True
env.mask = env.workspace + "\\final_mask.img"

# Check out the ArcGIS Spatial Analyst extension license
arcpy.CheckOutExtension("Spatial")

#---------------------------------------------------------------------------------
# Have user input the estimate values from the summary text file from Step 2.
## May include bathymetry, bathymetric relief, SST, distance from shore
## and/or any of the six sediment rasters.
intercept = arcpy.GetParameterAsText(1)
bathymetry_factor = arcpy.GetParameterAsText(2)
bathy_reli_factor = arcpy.GetParameterAsText(3)
sediments1_factor = arcpy.GetParameterAsText(4)
sediments2_factor = arcpy.GetParameterAsText(5)
sediments3_factor = arcpy.GetParameterAsText(6)
sediments4_factor = arcpy.GetParameterAsText(7)
sediments5_factor = arcpy.GetParameterAsText(8)
sediments6_factor = arcpy.GetParameterAsText(9)
dist_to_sh_factor = arcpy.GetParameterAsText(10)
Fall_2013_factor = arcpy.GetParameterAsText(11)
Spring_201_factor = arcpy.GetParameterAsText(12)

# Create the logit raster based on user inputs
inter = Raster(env.workspace + "\\final_mask.img") * float(intercept)
bathy = Raster(env.workspace + "\\bathymetry.img") * float(bathymetry_factor)
relief = Raster(env.workspace + "\\bathy_relief.img") * float(bathy_reli_factor)
seds1 = Raster(env.workspace + "\\sediments_1.img") * float(sediments1_factor)
seds2 = Raster(env.workspace + "\\sediments_2.img") * float(sediments2_factor)
seds3 = Raster(env.workspace + "\\sediments_3.img") * float(sediments3_factor)
seds4 = Raster(env.workspace + "\\sediments_4.img") * float(sediments4_factor)
seds5 = Raster(env.workspace + "\\sediments_5.img") * float(sediments5_factor)
seds6 = Raster(env.workspace + "\\sediments_6.img") * float(sediments6_factor)
dist = Raster(env.workspace + "\\dist_to_shore.img") * float(dist_to_sh_factor)
FSST = Raster(env.workspace + "\\Fall_2013_SST.img") * float(Fall_2013_factor)
SSST = Raster(env.workspace + "\\Spring_2014_SST.img") * float(Spring_201_factor)
```

```
logitRaster = inter + bathy + relief + seds1 + seds2 + seds3 + seds4 + seds5 + seds6 +
dist + FSST + SSST

# Convert to probability raster and save based on user selected file name
output_name = arcpy.GetParameterAsText(13)

exp_logit = Exp(logitRaster)
probRaster = (exp_logit)/(1 + exp_logit)

probRaster.save(env.workspace + "\\" + output_name + "_Prob.img")

#Convert to habitat raster using Youden-Index Cutoff
cutoff = arcpy.GetParameterAsText(14)

outCon = Con(Raster(output_name + "_Prob.img") >= float(cutoff),1,0)
outCon.save(env.workspace + "\\" + output_name + "_habitat.img")
```

Item 4. Final general additive model output for Scup. These results were not used in further analyses.



Item 5. Final general additive model output for Black Sea Bass. These results were not used in further analyses.



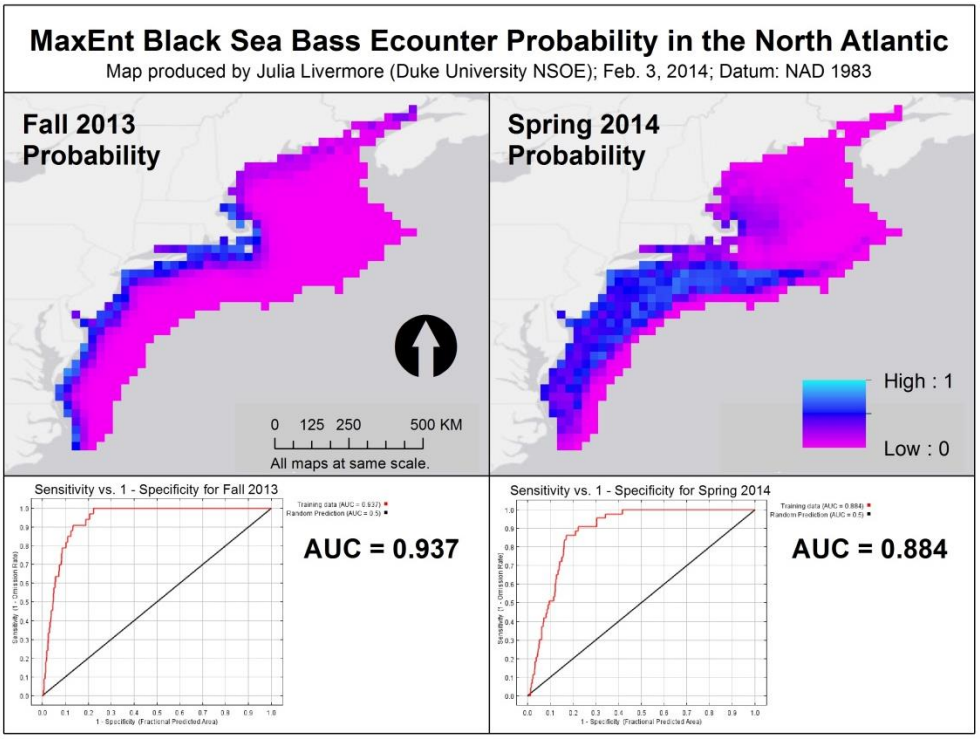
Item 6. Python script created to convert VTR data from an Excel file to a raster file in ArcGIS

```
##---------------------------------------------------------------------
## VTR_Data_Conversion.py
##
## Description: Read in VTR data from NOAA.
##
## Created: December 2014
## Author: Julia Livermore - jcl51@duke.edu (for master's research)
##---------------------------------------------------------------------

# Import system modules
import arcpy, os, sys
from arcpy import env
from arcpy.sa import *

# Set environmental settings
env.workspace = r'..\VTR_Data'
env.overwriteOutput = True

# Check out the ArcGIS Spatial Analyst extension license
arcpy.CheckOutExtension("Spatial")

#-------------------------------------------------------------------------------
# Set Local variables:
templateShp = env.workspace + "/template.shp"
```

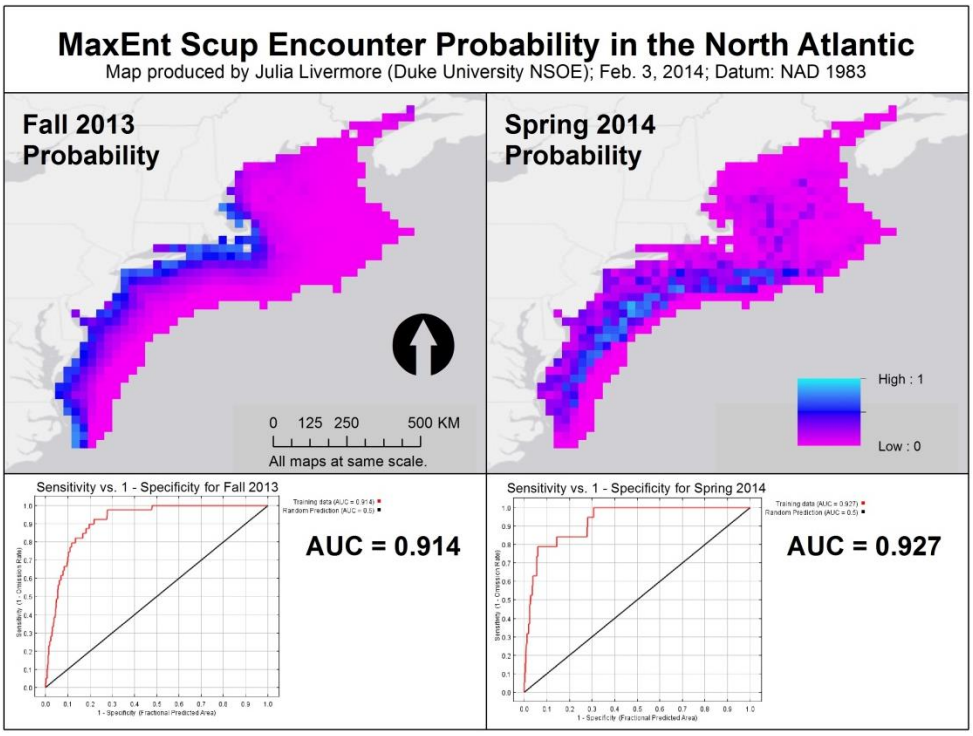
```
# Use Describe to get a SpatialReference object
spatial_reference = arcpy.Describe(templateShp).spatialReference

# Select input file
for root, dirs, files in os.walk(r"..\VTR_Data"):
    for file in files:
        if file.endswith(".csv"):

            # Create string for simpler paths
            SeasonSpecies = str(file[0:-4]) + "LBS"

            # Data Extraction from 1st NOAA VTR Data File
            print "Extracting " + SeasonSpecies + " data."

            # Create a shapefile of absence points from the survey data
            # Set Local variables:
            outputShapefile = SeasonSpecies + ".shp"

            # Create Feature Class
            arcpy.CreateFeatureclass_management(env.workspace, outputShapefile,
"POINT", "","DISABLED",
                                                "DISABLED", spatial_reference, "",
"0", "0", "0")

            # Add the points field
            arcpy.AddField_management(outputShapefile,"POUNDS","LONG")

            # Create an input cursor for the feature class so that we can add feature
records
            cur = arcpy.InsertCursor(outputShapefile)

            ## Open csv file for reading
            inputFileObj = open(file,'r')
            ## Start with first line and begin while loop through document.
            lineString = inputFileObj.readline()
            while lineString:
                # Parse line into a list
                lineData = lineString.split(',')
                # Extract attributes from the datum header line
                TNMS = lineData[0]
                Pounds = lineData[1]
                Boats = lineData[2]
                lastChar = int(TNMS[-1])
                lastChar2 = int(TNMS[-2])

                if lastChar == 1:
                    latMin = 55
                elif lastChar == 2:
                    latMin = 45
                elif lastChar == 3:
                    latMin = 35
                elif lastChar == 4:
                    latMin = 25
                elif lastChar == 5:
                    latMin = 15
                elif lastChar == 6:
                    latMin = 5
                else:
                    print "Error indexing ten-minute square latitude value.", lastChar

                if lastChar2 == 1:
                    longMin = 55
```

```
                elif lastChar2 == 2:
                    longMin = 45
                elif lastChar2 == 3:
                    longMin = 35
                elif lastChar2 == 4:
                    longMin = 25
                elif lastChar2 == 5:
                    latMin = 15
                elif lastChar2 == 6:
                    longMin = 5
                else:
                    print "Error indexing ten-minute square longitude value.",
lastChar2

                ddminlat = latMin/float(60)
                ddminlong = longMin/float(60)
                Lat = 1 * (int(TNMS[0:2]) + ddminlat)
                Long = -1 * (int(TNMS[2:4]) + ddminlong)

                try:
                    # Create a point object from the new feature class
                    obsPoint = arcpy.Point()
                    obsPoint.X = Long
                    obsPoint.Y = Lat

                    # Create a feature object to add to the feature class
                    featObj = cur.newRow()

                    # Set the feature's shape and other attribute values
                    featObj.shape = obsPoint
                    featObj.setValue("POUNDS",Pounds)

                    # Commit the feature to the feature class
                    cur.insertRow(featObj)
                except Exception as e:
                    print e, "Error adding ten-min square value " + TNMS + " to the
file."

                # Move to the next line to continute the while loop.
                lineString = inputFileObj.readline()

            # Close the file object and delete cursor
            inputFileObj.close()
            del cur
```

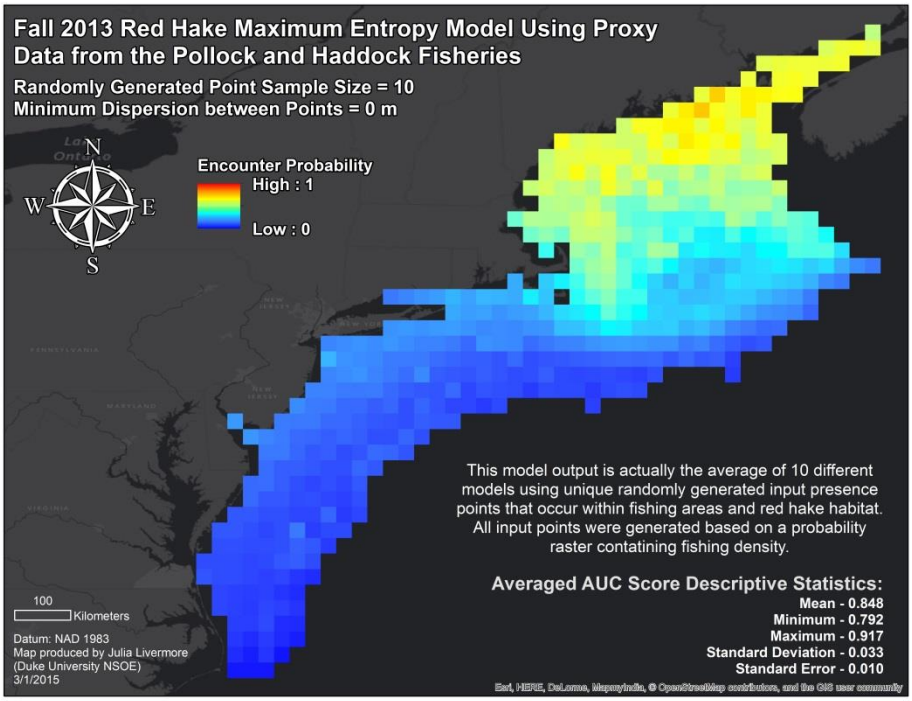
Item 7. Final maximum entropy model output for Black Sea Bass using presence data from the NEFSC trawl surveys. These results were not used in further analyses.



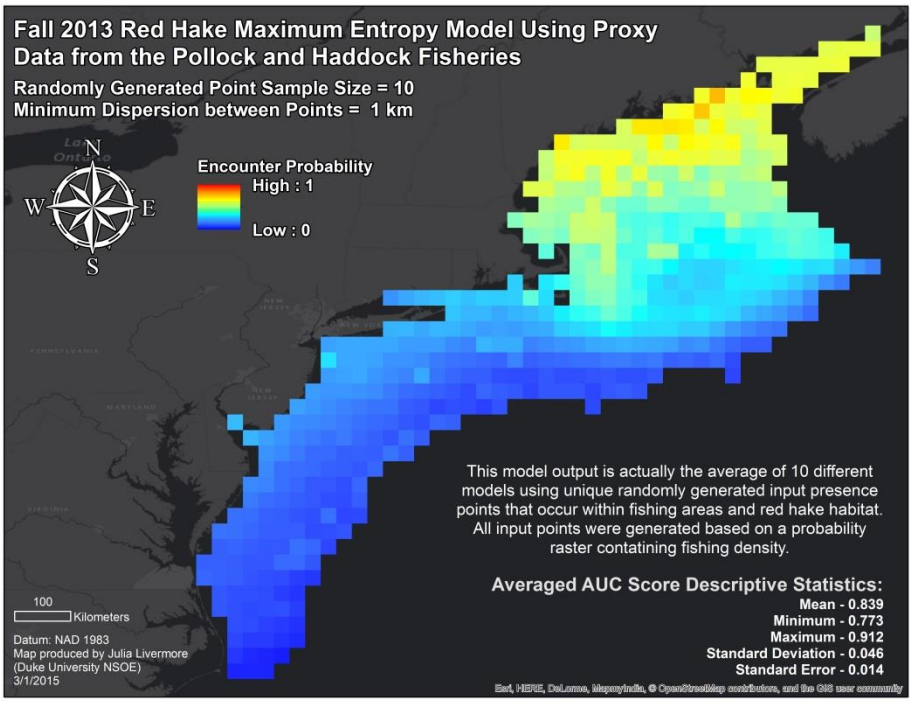
Item 8. Final maximum entropy model output for Scup using presence data from the NEFSC trawl surveys. These results were not used in further analyses.



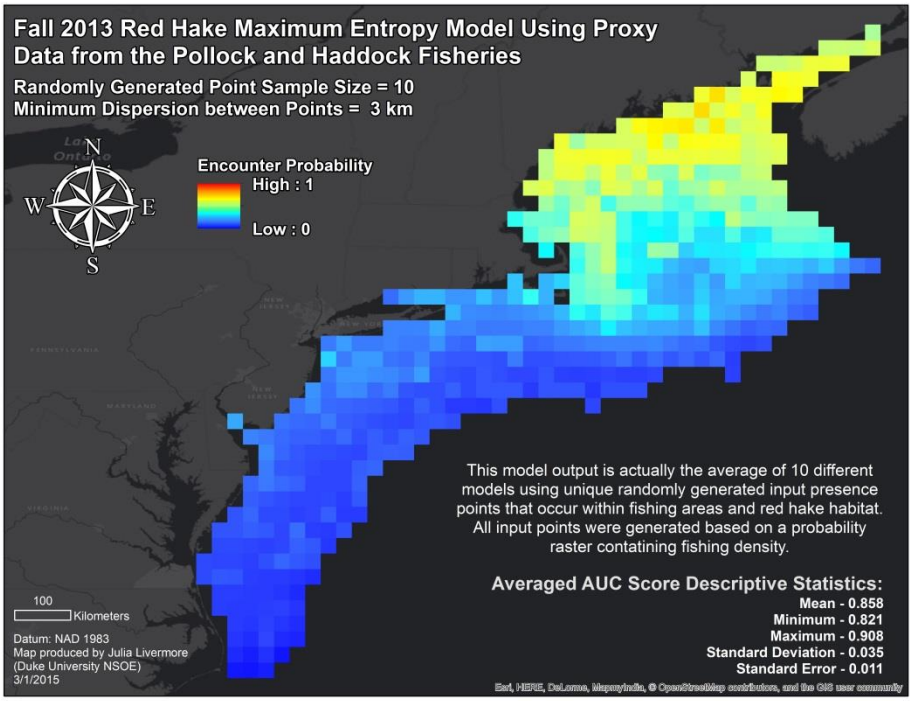
Item 9. Final averaged maximum entropy model for Red Hake in fall 2013 using 10 uniquely and randomly generated input point datasets with 10 points and 0 km minimum distance between points.



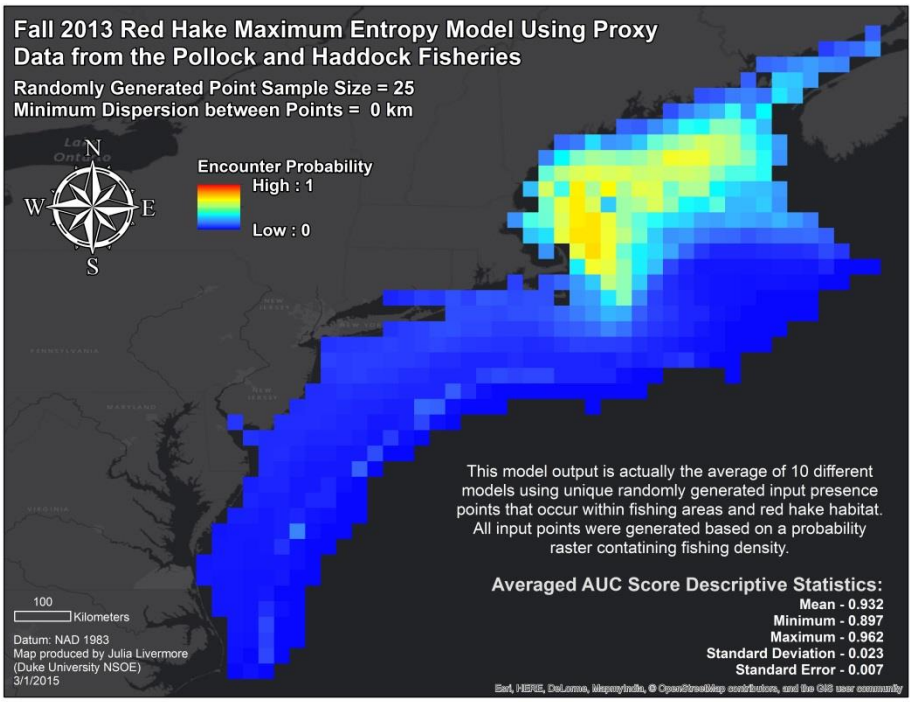
Item 10. Final averaged maximum entropy model for Red Hake in fall 2013 using 10 uniquely and randomly generated input point datasets with 10 points and 1 km minimum distance between points.



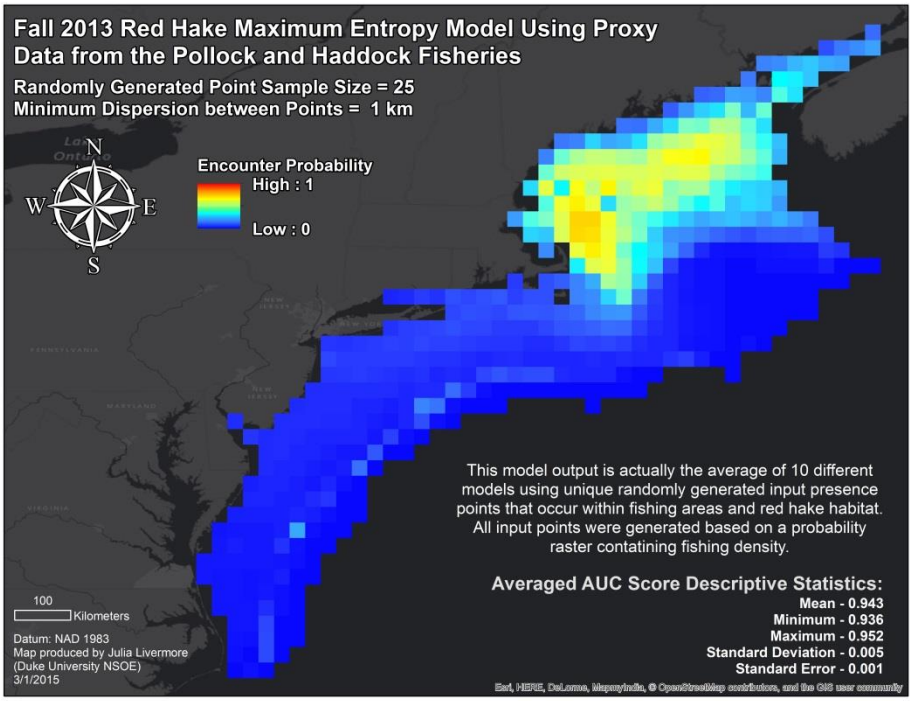
Item 11. Final averaged maximum entropy model for Red Hake in fall 2013 using 10 uniquely and randomly generated input point datasets with 25 points and 3 km minimum distance between points.



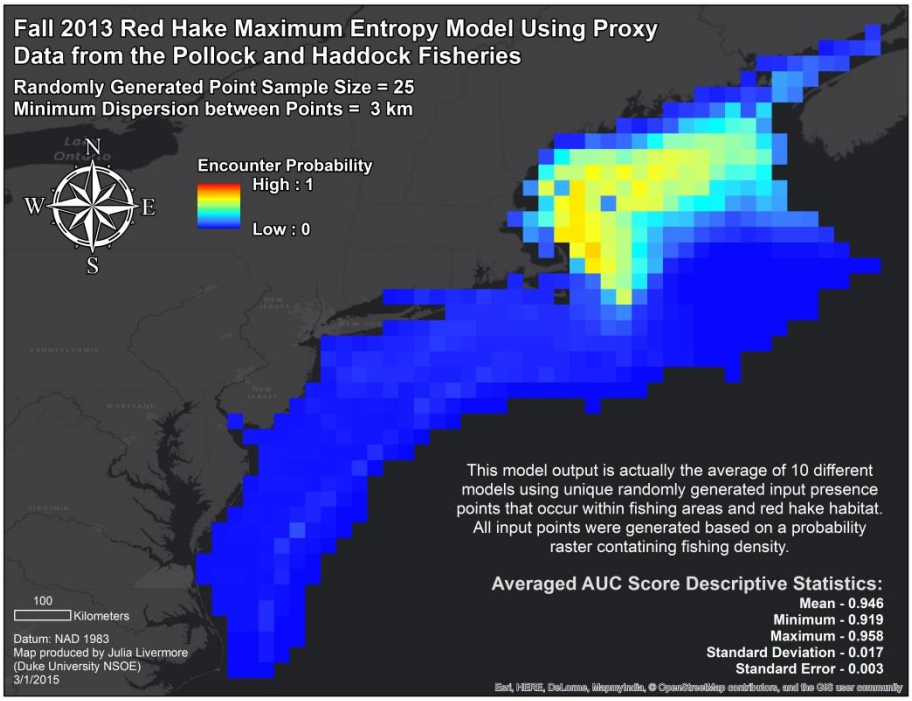
Item 12. Final averaged maximum entropy model for Red Hake in fall 2013 using 10 uniquely and randomly generated input point datasets with 25 points and 0 km minimum distance between points.



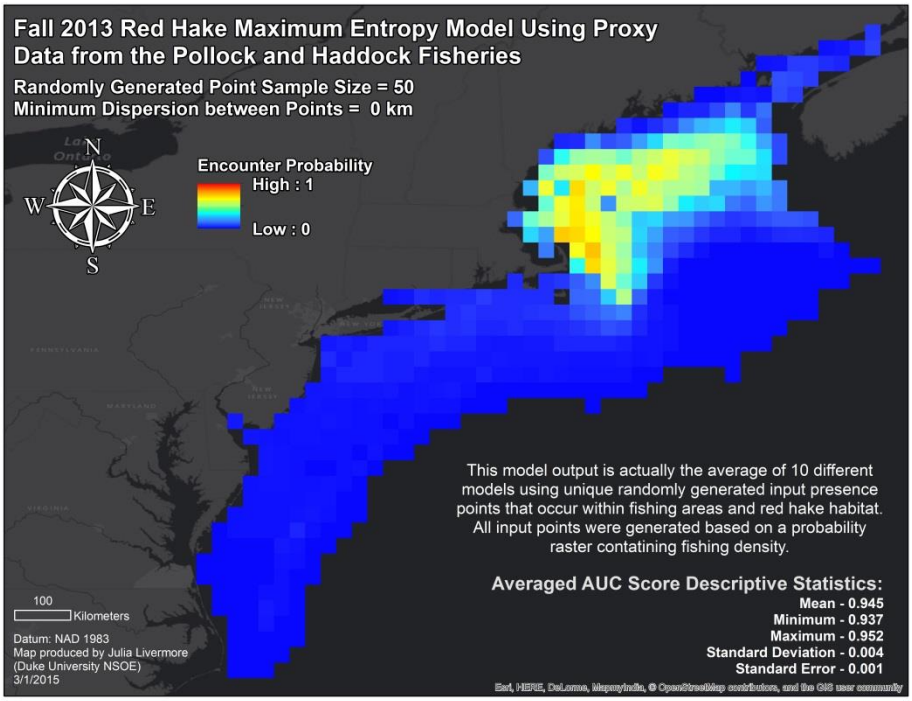
Item 13. Final averaged maximum entropy model for Red Hake in fall 2013 using 10 uniquely and randomly generated input point datasets with 25 points and 1 km minimum distance between points.



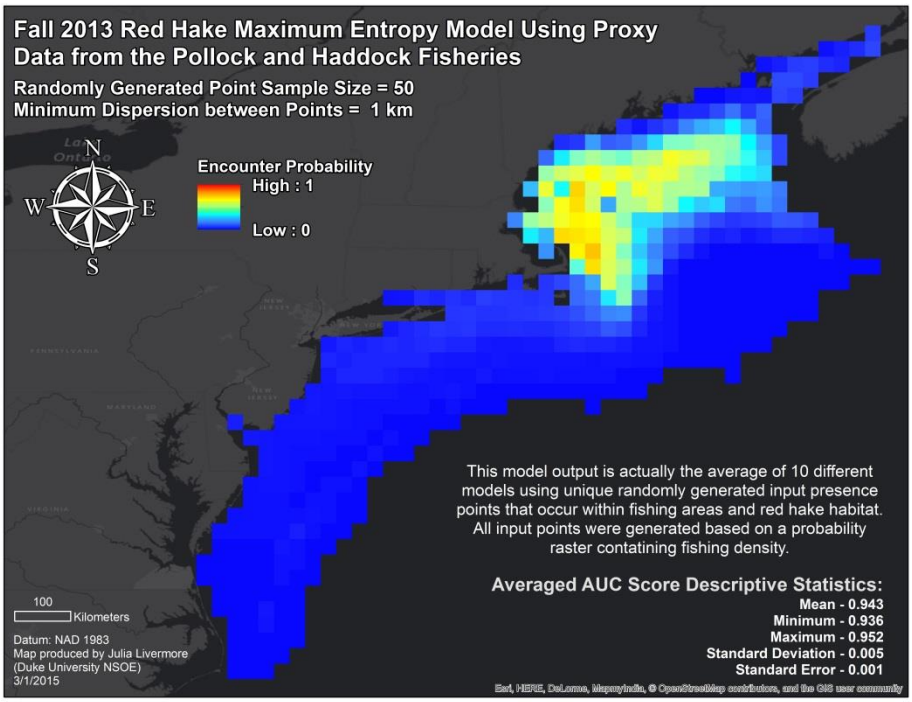
Item 14. Final averaged maximum entropy model for Red Hake in fall 2013 using 10 uniquely and randomly generated input point datasets with 10 points and 3 km minimum distance between points.



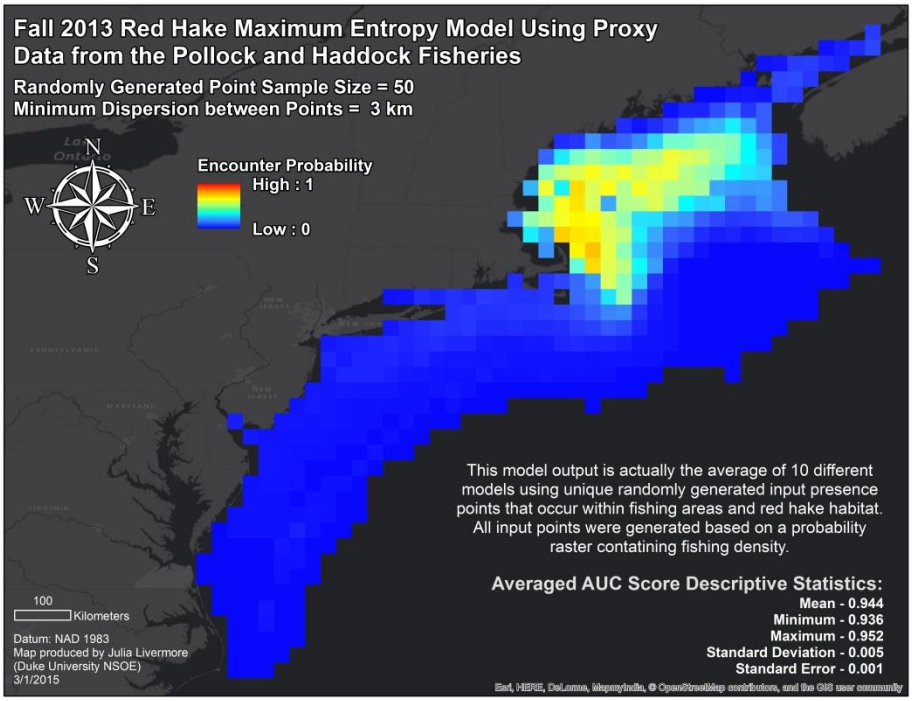
Item 15. Final averaged maximum entropy model for Red Hake in fall 2013 using 10 uniquely and randomly generated input point datasets with 50 points and 0 km minimum distance between points.



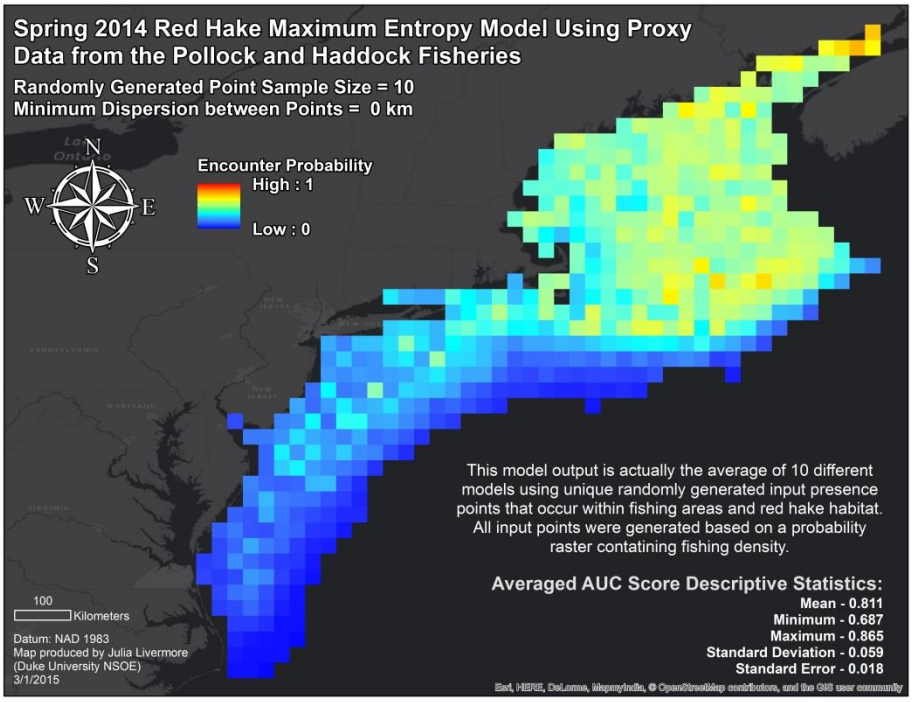
Item 16. Final averaged maximum entropy model for Red Hake in fall 2013 using 10 uniquely and randomly generated input point datasets with 50 points and 1 km minimum distance between points.



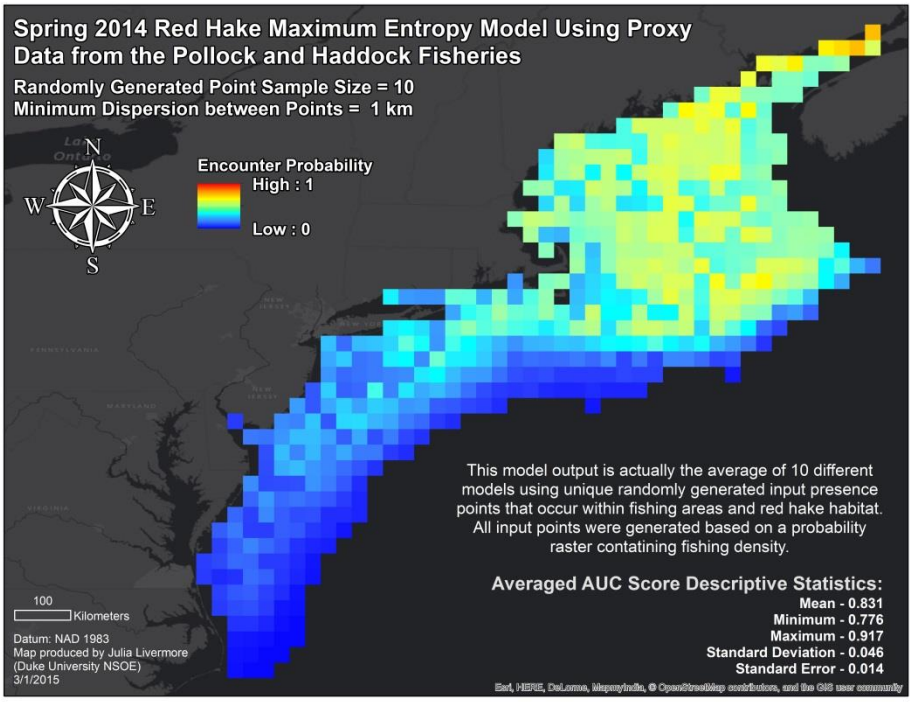
Item 17. Final averaged maximum entropy model for Red Hake in fall 2013 using 10 uniquely and randomly generated input point datasets with 50 points and 3 km minimum distance between points.



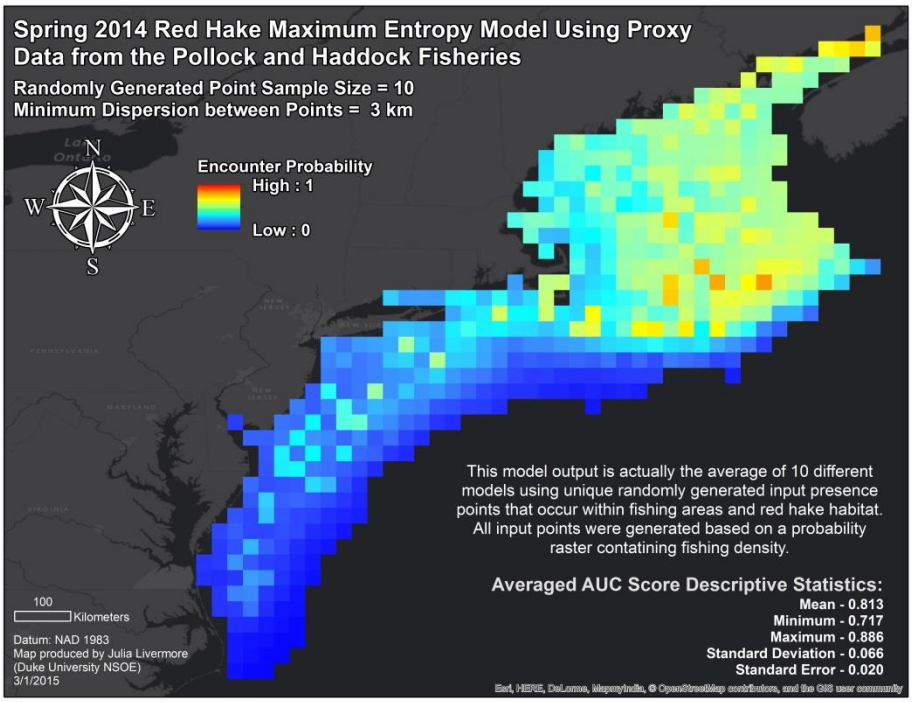
Item 18. Final averaged maximum entropy model for Red Hake in spring 2014 using 10 uniquely and randomly generated input point datasets with 10 points and 0 km minimum distance between points.



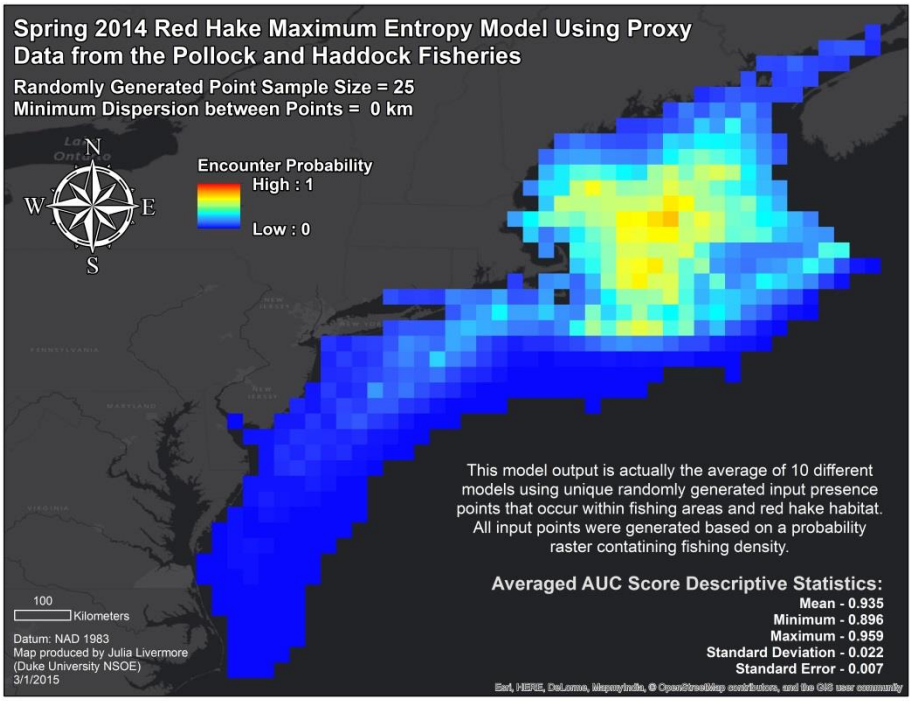
Item 19. Final averaged maximum entropy model for Red Hake in spring 2014 using 10 uniquely and randomly generated input point datasets with 10 points and 1 km minimum distance between points.



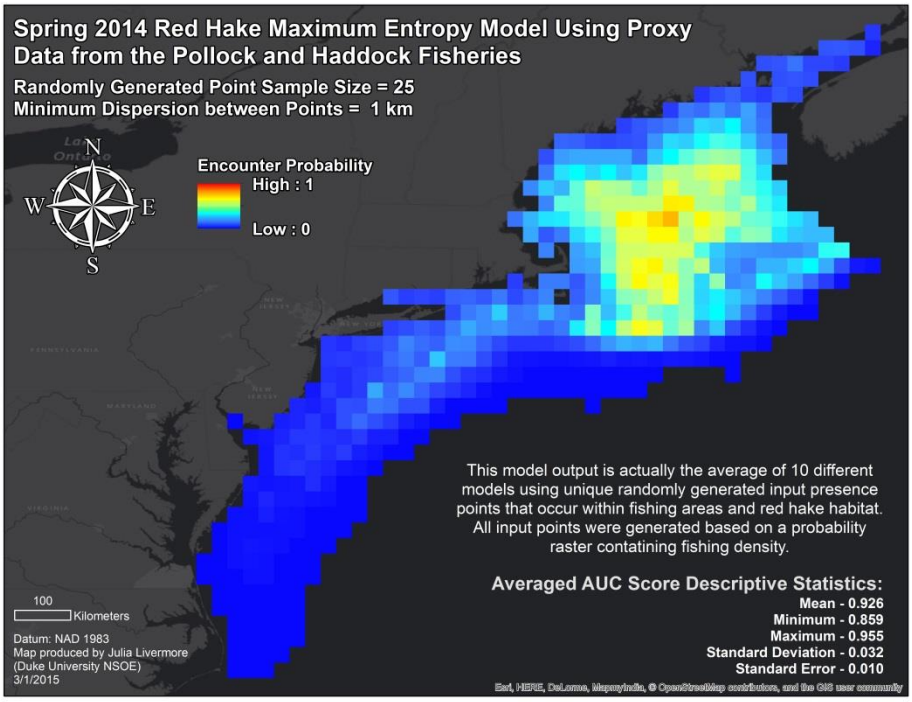
Item 20. Final averaged maximum entropy model for Red Hake in spring 2014 using 10 uniquely and randomly generated input point datasets with 25 points and 3 km minimum distance between points.



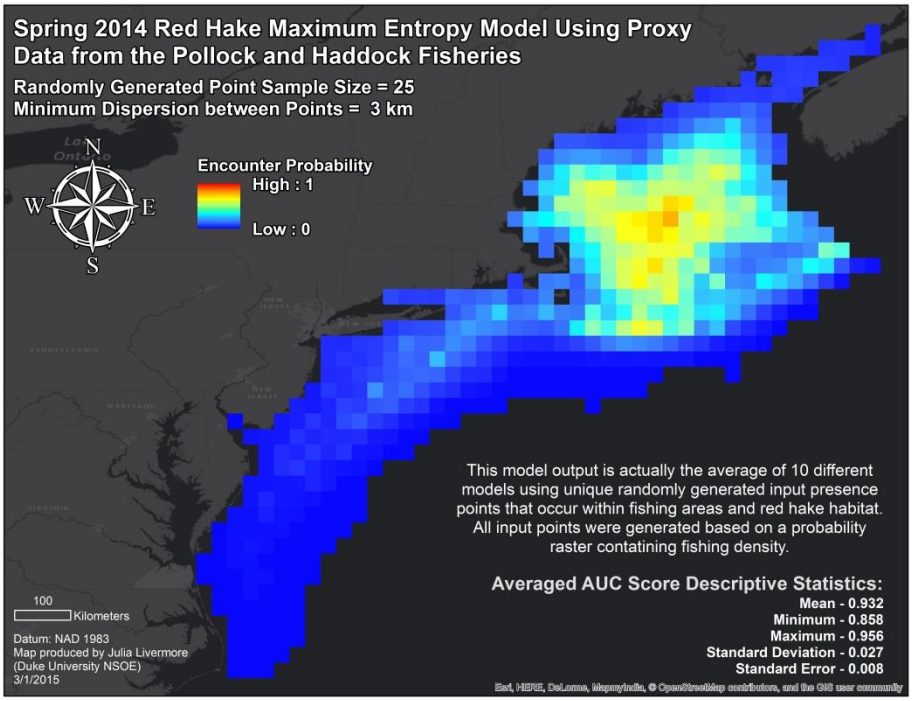
Item 21. Final averaged maximum entropy model for Red Hake in spring 2014 using 10 uniquely and randomly generated input point datasets with 25 points and 0 km minimum distance between points.



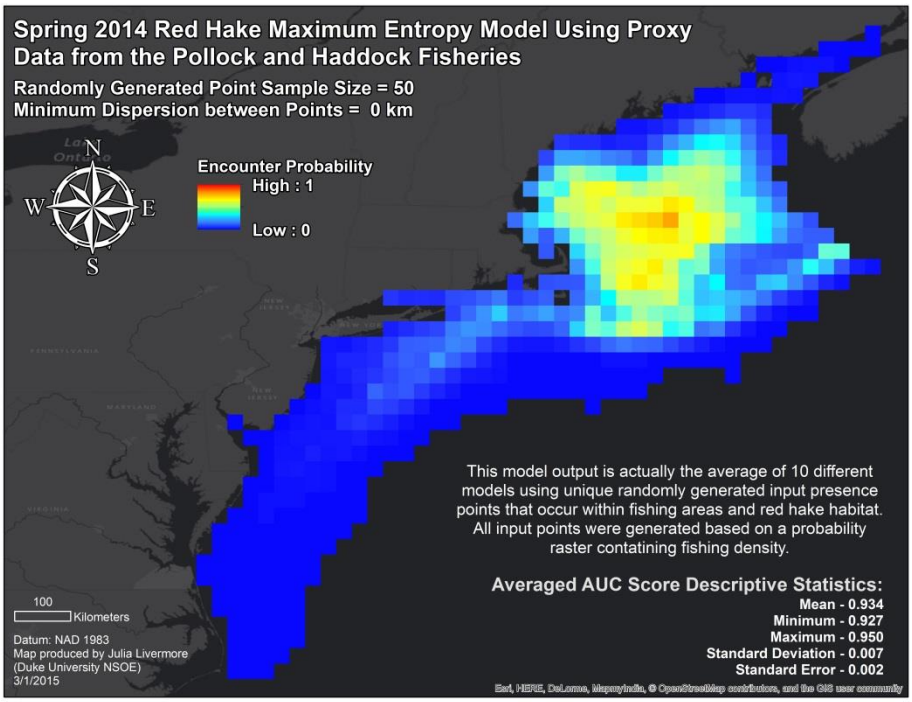
Item 22. Final averaged maximum entropy model for Red Hake in spring 2014 using 10 uniquely and randomly generated input point datasets with 25 points and 1 km minimum distance between points.



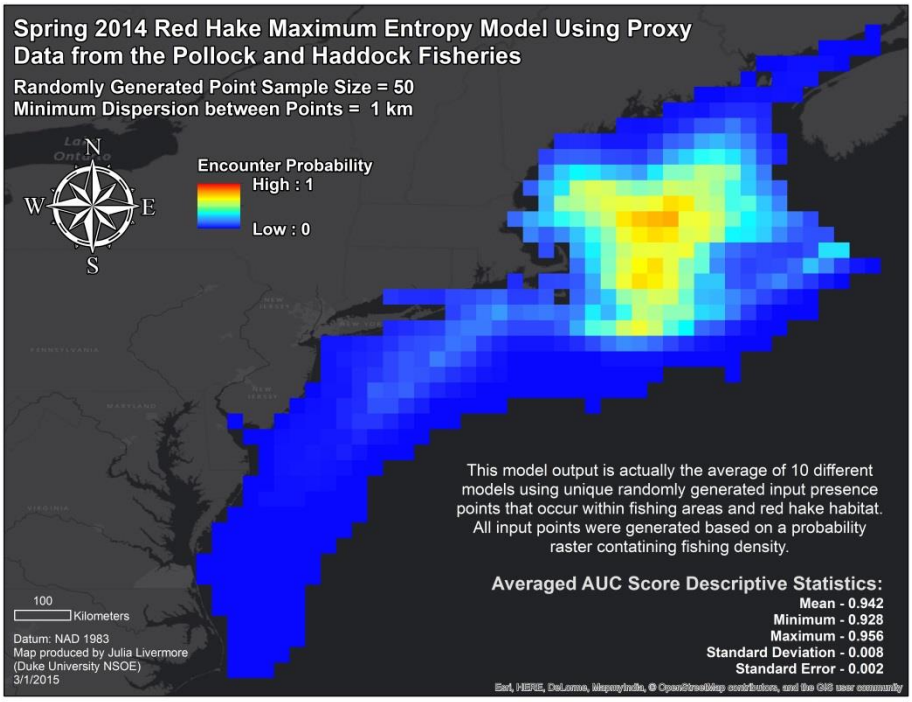
Item 23. Final averaged maximum entropy model for Red Hake in spring 2014 using 10 uniquely and randomly generated input point datasets with 10 points and 3 km minimum distance between points.



Item 24. Final averaged maximum entropy model for Red Hake in spring 2014 using 10 uniquely and randomly generated input point datasets with 50 points and 0 km minimum distance between points.



Item 25. Final averaged maximum entropy model for Red Hake in spring 2014 using 10 uniquely and randomly generated input point datasets with 50 points and 1 km minimum distance between points.



Item 26. Final averaged maximum entropy model for Red Hake in spring 2014 using 10 uniquely and randomly generated input point datasets with 50 points and 3 km minimum distance between points.